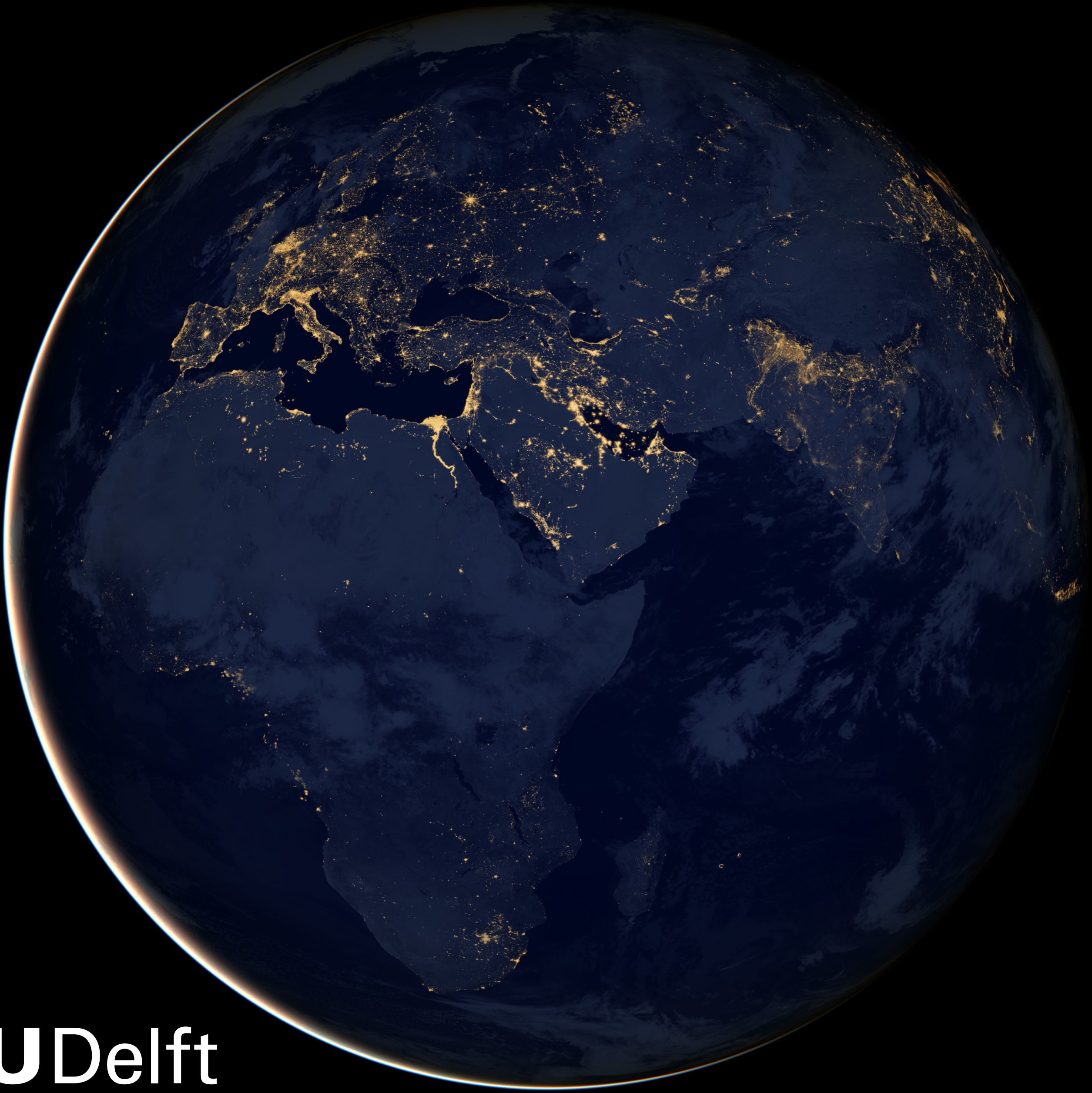
Fairness Information Maximization on Social Media

Course name: IN5000

Distributed algorithm Group

Delft University of Technology



Author:	Zhiyue Zhang
Supervisor:	Lydia Chen Ya-Wen Teng
Course code:	IN5000
Course period:	Quarter one to Quarter four
Acedemic yeaer:	2020-2022

The last update of this report has been made on Thursday 31st March, 2022.

Abstract

The rapid growth of the Internet use has allowed social networks to become the most effective means for marketing, leading to the emergence of "viral marketing" as a business model. The biggest challenge that is facing "viral marketing" is selecting seed users from the whole user set to form a "seed-set" to spread the influence and maximize the number of influenced users. This is known as the classic influence maximization problem. Based on the background provided, the paper focus on the fairness of information Maximization in social media and try to explain how the homophily effect, rich get richer mechanism and the duration of top users impact users in temporal social network. The paper also aims at developing a time-Awareness disparity seeding framework based on Disparity seeding framework, which is proved by experiments to slove the absolute error problem that exist between the target ratio and influential ratio. Furthermore, the unequal seeding disperse algorithm (USD), equal seeding disperse algorithm(ESD) and origin seeding disperse algorithm(OSD) have been developed to improve the influence maximization in temporal social network. The purpose is to find the most cost-effective user seed-set in any given period of time to maximize the influence of target users and increase the number of influenced users. According to the experimental test result, it is found that unequal seeding disperse algorithm perform better than the other two algorithms. In this paper, many experiments are carried out to verify the effectiveness of all the three algorithms using real social network data sets. As a result, the effectiveness and efficiency of the proposed algorithm was proven.

Preface

This thesis concludes an incredible three-year journey to pursue a Master of Science in Computer Science from TU Delft. Personally, the road has a lot of ups and downs, and I wouldn't be able to handle them without Ayoub's help, for which I am grateful. I'd want to express my gratitude to my family, particularly my parents, for their unwavering support and endless prayers for my master's thesis research, also to my friends, who never fail to brighten my day.

Throughout completing this thesis, I received a lot of help from a lot of people. First and foremost, I'd like to express my gratitude to Prof. Lydia Chen for working on this thesis under her supervision. Lydia is always willing to help with research and allows me a lot of leeways to concentrate on the topics that interest me the most. Every time we discuss the development of my thesis, her critical yet motivational comments have helped me shape my research mentality.

I'd also like to express my gratitude to Dr. Ya-Wen Teng for providing me with the incredible opportunity to conduct research with NCSIST. Ya-Wen has always been available for talks and has provided extensive critical input on the outcomes of my experiment and the drafting of this thesis regularly.

I also want to thank my classmates at TU delft (Maciej) and Ting Wei Li from National Taiwan University for our insightful discussions every time I presented my work online. Thank you for sharing all of the great moments with me as we tackled this meaningful journey.

Zhiyue Zhang
Delft, March 2022

List of Figures

List of Tables

Contents

1 Introduction

This paper describes the primary research significance, research methods of tribute and discusses the organizational structure of this paper. This chapter first summarizes the phenomenon of influence force propagation in social networks and introduces the research hotspots initiated by this phenomenon. Then, it introduces this paper's background knowledge from the basic concept of social networks research status, research trends, and difficulties.

1.1. Motivation

The social life is built upon people's interaction. A social network is a network of real-life people and their social relationships. Reflecting the connection and communication between people, the rapid development of the Internet has broken the time and space restrictions of offline social activities, and people can carry out rich social interaction online anytime and anywhere. Build and maintain your social network and interpersonal relationships.

With the rapid advancement of information technology, online social networks such as QQ and Facebook are becoming increasingly popular [29]. In social networks, ideas and other information can be quickly spread to more users through reading, forwarding, and recommending by each user. This is the phenomenon of influence force transmission in social networks. There are many famous examples of influence force transmission in social networks in real life. The best example is the ALS ice Bucket fight, popular in 2014. Only a few users uploaded the challenge video on Facebook and then sent it to their friends to ask them to accept the challenge. After a period of spread on the network, more than 2.4 million people uploaded the challenge video on Facebook. And raised $982 million to help with the treatment and care of ALS patients. Such is the power of influence spread through social networks.

It offers fantastic potential for promoting new items or disseminating political agendas in a short amount of time. To develop a new film, for example, select some influential users from social networks, then offer them a prize and urge them to utilize the Word-of-Mouth effect in social networks to excite the attention of a large number of people [36]. When implementing viral marketing, we normally strive to discover people with a lot of clout who can disseminate content to a large number of people. Because of this, how does the communication of social network influence can be better used by people, which has become a hot research topic.

The basic problem of Influence maximization is the Influence maximization question, which is short for IM question. Given constant K, it is required to select K users in the social network to construct a seedset. In addition, users are activated by the influence force propagated by seedset to maximize the expected number of users in the active state in the social network.

However, resolving the influence maximization problem, which seeks to identify the most influential individuals, poses specific challenges [14]. The study of identifying a small group of influential people in a social network in order to maximize their influence [25].

However, there are some systematic gaps in traditional IM algorithm problems. Firstly, in real life, the transmission of influence force has certain timeliness. Propagation within a specific time interval is practical, not the norm of the infinite time interval. Secondly, the consumption cost of different users being selected into the specific subset and the feedback benefit brought by being affected should be different. This is determined by the importance of different users in the social network and other properties. For example, each influential user in the network is significantly different due to its fans. As a result, when businesses care about promoting products through viral marketing, The fees of big influential users are much higher than those of small influential users, and the final marketing effects are also very different, bringing different benefits to businesses.

Moreover, through communication, society could solve labor division between men and women, as well as the proportion of the population. However, there is a sizeable male-female disparity at several levels in today's society also uneven proportion of men and women who occupy different existing jobs. According to researchers [41], there is a correlation between the (perceived) gender of users, gaining visibility and influence on social media. This proves the existence of the glass ceiling effect, which makes it harder for females to become highly influential. However, users with higher visibility are frequently chosen as seeds as a specific group for dissemination, ensuring maximum dissemination efficiency becomes an interesting topic. The aim has historically been to increase the total number of people affected in the process by identifying a collection of initial sources (i.e., seed nodes) in a social network who can influence other people (e.g., who received the information being propagated) [2]. But ensuring that women's visibility on social media is fair is a critical issue.

Therefore, based on the above discuss, this study considers the time constraint factors and conducts differentiated method on the effectiveness and social influence maximization of different graphs in social networks. In this case, the first question studied in this paper is the difference between users' behaviour pattern in the time-sensitive social network graph and the traditional static graph. Furthermore, solve the problem through quantitative analysis. On this basis, This paper presents a time-sensitive social network communication influence maximization question. The purpose of this paper is to select seed sets that meet budget constraints to maximize social influence within time constraints. Then, the gender equity problem of SeedSet Users is further studied on the time-sensitive social influence diffusion model. This paper designs a time-awareness Disparity seeding algorithm to solve this problem. Experiments on real data sets verify the performance of the algorithm.

The chapter starts with discussing the problem definition, followed by the research question and several associated sub-questions. Literature review have been used to acquire the requirement needed for designing the algorithm and discovering the research question. Moreover, in the last part of this chapter, we discuss the research contributions.

1.2. Problem definition

Based on our preliminary understanding of the research on the propagation of the influence force of social networks and the literature study that has been performed for this thesis, some important challenges are discussed as follows. These problems eventually become the strong motivation for developing research ideas and conducting experimental experiments.

1.2.1. The temporal effect of influence transmission in social networks

For the research on the propagation of the influence force of social networks, the social influence model is constantly developing to meet the requirement of reality, and the algorithm is increasingly not satisfied with the simple heuristic. For this reason, the extension of social influence diffusion model and algorithm optimization are the development trends of social network influence propagation research. The studies on temporal factors in the transmission of the social influence of social networks are not complete enough, and the relevant questions proposed are not sufficiently relevant to the actual situation. Considering the role of temporal factors and user differences in the transmission of interaction force of social networks comprehensively, this research aims to propose a more fundamental problem of the social influence propagation and give a fast and effective algorithm.

1.2.2. Influence maximization of the social network

Based on the in-depth investigation of the research status, this paper analyzes how to try to reach influence maximization of the propagation of the target seed sets and how to estimate the characteristics of graphs in different periods in the process of this research. Furthermore how to design gender fairness strategies based on influence maximization. The above three questions are the main questions to be solved in this study.

The calculation of influence transmission is the most crucial issue in researching influence force transmission in social networks. However, the near-similarity algorithm for calculating feedback benefits of influence force transmission under traditional models is no longer applicable in time-sensitive communication models. A strict mathematical analysis is needed for the limitation of temporal factors.

1.2.3. Gender fairness in the social network

Recent research [16, 18, 19] has focused on reducing the gender gap while increasing information influence force. However, they cannot assure that the influenced users have a gender ratio to meet specified demand. In the paper [52] they proposed disparity seeding, which goes one step further to maximize overall information transmission and achieve any desired gender ratio by design. They propose to use a influenced ratio to minus target ratio to get the absolute error to verify if the disparity seeding framework meet their demand. However, there is no too much research about how to reduce the absolute error between influenced ratio and target ratio in dynamic network graph while considering temporal effect.

1.3. Research question

Based on the challenges that have been identified in Section 1.2, we formulate three research questions which we will address by conducting thorough analysis on the dataset from Facebook and INS.

Research Question 1. *Can we identify the behavior patterns and characteristics of users in temporal social networks?* Related to the first challenge we have described earlier, we begin our investigation by defining the problem we want to address. In particular, we focus on the challenge of user behavior, given past interactions of a user for a temporal social network, and Summarize user behavior patterns and norms. Based on [46] People's interests vary throughout time as a natural process, especially in situations where they regularly engage with a wide range of goods. At the same time, as new nodes and edges indicate new interactions/relations in the underlying social structure, these social networks expand and change rapidly over time. Therefore, we believe that users' behaviors in the temporal social network have more realistic value and research significance because time flows rather than static in real life, and people's behaviors will change with time. Note that in a dynamic social network, the behavior between users and the interaction mode between users who with different influence levels, and whether the interaction mode of users will change over time.

According to [46] new links and nodes are constantly produced across social networks as new people join the network, and new friendships are formed, just as preferences vary over time. This leads to various functional analyses, including event and anomaly identification, in evolutionary networks analysis field. Overall, it is important to highlight that the user behaviour in temporal social network, Ours focuses on analyzing the behavior patterns and characteristics of users in temporal social network, and propose a novel result about this question. It is important to choose different strategies based on user characteristics at different times. We are particularly interested in unequally allocate seeds in different time period and consider to use unequally seeding algorithm to do the seed allocations.

Research Question 2. *Can we improve the disparity seeding algorithm by accommodating temporal effect in social networks to reach fairness of gender ratio for seeds selection?* Related to the challenge of the dynamic distribution method of seeds set In temporal social network, given past interactions of a user for a particular graph size and the characteristic, we propose an time-awareness disparity seeding algorithm which leverages this information. In particular, we are interested in checking whether can dynamically adjust the size of seed sets based on graph size and graph inner intensity, in order to reach the goal of fairness of gender selection. Lastly, we also verified target HI-index is the best index among the other indexes for reach the goal of gender fairness. We perform analysis on the change of graph size over time and compare the result with the baseline method.

Research Question 3. *Can we improve the disparity seeding algorithm by accommodating temporal effect in social networks to achieve the goal of influence maximization?* Concerning the challenge of leveraging multiple factor and multi-dimensional information, we explore the different possible ways to incorporate these elements into the time-awareness disparity seeding algorithm and diffusion algorithm. We tested out various time period and variables related to the challenge of how to reach the influence maximization by using disparity seeding algorithm. In particular, we are interested in The relationship between the spread of social influence and time factor. Because it is unclear whether the seed users will have the similar influence pattern in when they are in static graph or not.

It is important to note that we will not predict the user patterns. Our focus is to verify whether seed users will have a similar pattern and the characteristics of users' dynamic behavior patterns. We validate the performance with different analysis methods for our designed algorithms.

1.4. Contributions

The main contribution of this paper is reflected in the following three aspects:

1. **Data analysis of organic User behavior and influence dissemination pattern in temporal social network.** Because temporal aspects have been overlooked in research on the impact of social network social influence transmission, we conducted characteristic analysis and behavior pattern analysis for different user groups based on big data statistical analysis.
2. **The design of time-awareness disparity seeding algorithm for fairness of seeds gender ratio and influence maximization.** In the communication of social networks, different genders have different communication influences. This study hopes to find a seed set that can meet the fairness of gender ratio under time-constrained conditions. In the communication of the influence force of social networks, different users will bring different communication benefits after being influenced by the influence force. Aiming at the time-sensitive influence force propagation model, this study hopes to satisfy the given time constraints. Find a seed set that maximizes the feedback benefit of all active users when the time constraint is reached.
3. **Insights into social influence diffusion pattern in temporal graph.** We get a beneficial conclusion by analyzing the experimental results, which can help us have a more precise and more quantifiable cognition of the user's interactive behavior in the dynamic graph.

1.5. Research methodology

For the user characteristics in time-awareness social networks, the user behavior is firstly studied based on the correlation theory of mathematical statistics. For the time-awareness influence propagation maximization problem, we refer IM algorithm and previous paper algorithm [52]. Then, the time factor and dynamic graph structure factor are combined into the algorithm. Furthermore, based on the influence propagation maximization algorithm, the paper optimizes the disparity seeding algorithm then develop the time-awareness disparity seeding algorithm, in order to achieve our goal of fairness of gender ratio and also reach the influence maximization.

The research design mainly based on the three stages:

1. **Discover and understand** defining the goal of the thesis and understanding the barriers from literature review.
2. **Design and analysis** designing a prototype based on the goal and analysis the result of the experiment.
3. **Conclusion** Forming the conclusion based on evaluation.

This research design follows three stages, the first stage(understand and discover) includes research ideas, literature review, research group discussion, and the theoretical formulation of the research questions. The first step of the research that has to be taken is to have a research idea and conduct a literature survey that revolves around the research idea and, at the end of the stage, finishes the theoretical formulation of the research questions. The knowledge of prototype design can be obtained from scenario analysis and literature review.

The first step of the research that has to be taken is to have a research idea and conduct a literature survey that revolves around the research idea, and after that we perform several research group discussion to figure out what is the current problem we need to solve. At the end of the stage, finishes the theoretical formulation of the research questions.

On the quantitative empirical research stage 2, the knowledge of temporal effect analysis can be obtained from data analysis and literature review. The design time-awareness disparity seeding algorithm and diffusion algorithm are based on the result of data analysis. The update and both algorithm can be improved through communication with research group, the absolute error of the final result and the influence maximization result.

On the third stage, once the experiment ends, we analyze these results and try to get some useful conclusions from that.

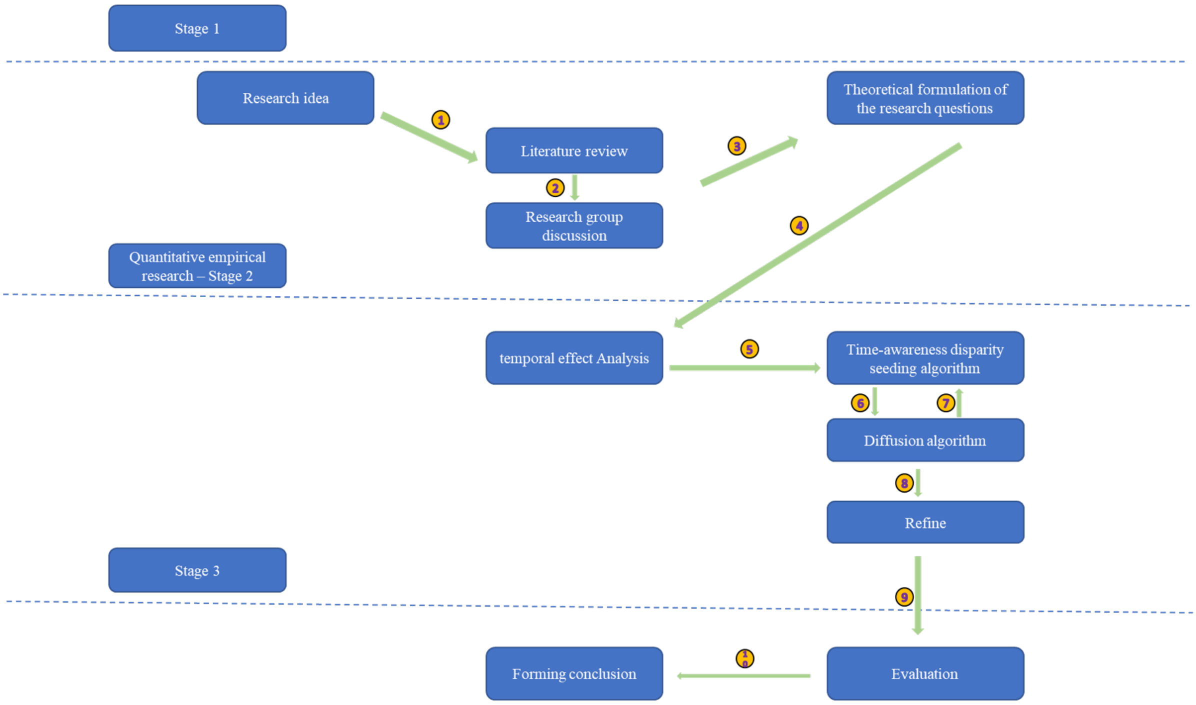

Stage 1
Research idea
1
Literature review
2
Research group discussion
3
Theoretical formulation of the research questions
4
Quantitative empirical research – Stage 2
temporal effect Analysis
5
Time-awareness disparity seeding algorithm
6
7
Diffusion algorithm
8
Refine
Stage 3
9
Evaluation
10
Forming conclusion

Figure 1.1: Research approach

1.6. Research thesis structure

This paper is divided into six chapters. The first chapter is the introduction. It first introduces the phenomenon of influence propagation in social networks and illustrates the characteristics of user behavior generated by it. Step forward, based on some basic assumptions, give the research questions of this paper. Next, discuss the state of arts academic research on these questions and analyze the research's development trend and difficulties based on the above analysis. At the end of this chapter, this paper's research significance and research methods are described. The second chapter is the literature review. First of all, the research on social the interaction force of social networks is carried out, and the influence of time on user behavior is discussed. Then, several information influence models are introduced. Finally, the application of time awareness model is introduced. The Third chapter is the temporal analysis. This chapter analyzes the time factor for user behavior and answers one of our research questions. Next, we analyze three important problems affecting user behavior and conclude that. The forth chapter is the algorithm. We first introduced different metrics for measuring the intensity of information influence, followed by our Framework for the Disparity seeding algorithm and then three different algorithms. The fifth chapter is the implementation of the algorithm and data analysis.Finally, we describe the conclusions recommendations and future work in Chapter six.

2 Literature review

Our research touches on a variety of areas, including user behavior, social networks, and temporal data analysis in general. This chapter identifies, organizes, and discusses the current state of the art. Our proposal's uniqueness rests at the intersection of these two areas.

The broad concept of the influence maximization method and different specific techniques and algorithms are covered in this chapter. First, in Section 2.1, We introduce the temporal dynamics of user behaviour. And then, in Section 2.2, We introduce the temporal social network. In section 2.3, discusses the state-of-the-art algorithms in influence maximization algorithm. Special attention is given to IM models called time-aware diffusion models, which we are going to use extensively in our approaches later. Section 2.4 explains the time aware IM application.

2.1. Temporal dynamics of user behaviour

In this section, we conduct a static and dynamic characterization of female and male interaction patterns review, try to explore the results of research on homophily effect, Glass ceiling effect and other user behaviours in temporal effect.

2.1.1. The study of homophily effect

Homophily prefers interactions with people who are similar to them, and it has been shown to hold for social networks and online communities [15, 38, 50]. The most recent showing that rich-get-richer dynamics mixed with homophily naturally increase the advantage of a majority group is particularly applicable to our situation [4]. That approach can explain why organic engagement on Instagram, like other platforms, exhibits the bias indicated above against some of its users. However, this paper does not consider the influence of the time factor on the rich get richer phenomenon.

[60] developed a theoretical framework for understanding sociodemographic structures that influence various tie formation mechanisms and analyze the racial homogeneity of networks, furthermore prove same-race friendships are likely to develop in mentioned aggregation effects implied by ethnic homophily. However, there is not much contribution to developing an approach that would help determine why certain principles of homophily relationship formation take precedence over others in certain circumstances. These principles are causally interrelated once their effects have been disentangled and how they interact to produce longitudinal dynamics as a social network graph change. based on the [8] research, they used log-multiplicative family of models to analysis, which showing the trend toward homophily and social distance for males and females in 1985 and 2004, finding homophily and social space is comparatively stable. And Time is essential for changing homophily and social space for males and females. Education did not play such a vital role in that.

2.1.2. User behaviour in temporal effect

According to [63]'s research, they propose a continuous temporal approach, a continuous temporal dynamic behavior model, and then based on the previous method, to get the generation of temporal behaviors. However, they are missing how social influence changes with the temporal behavior effect change. [27] first time to use embedding to solve the influence maximization problem. They propose the adversarial graph embedding method and form a suitable selection method. However they more focus on the minority group's fairness information spreading did not focus on gender.

2.1.3. Glass ceiling effect

Our research addresses recent concerns about the fairness of decisions based on Big Data algorithms. Such algorithms can potentially establish unfair prejudices even if there is no express aim to discriminate and without access to sensitive factors such as gender and ethnicity [6]. On various online social platforms, several research look into the glass ceiling effect. Male users, on average, acquire greater visibility and share information more quickly [50, 31, 41].[12] mentioned that the majority of this issue's associated research focuses on binary classification jobs like loan approval to maintain operational efficiency under diverse conditions of equal treatment or opportunity. [7, 62] investigated fairness for word embeddings and node ranking. [50] found that homophily and growth dynamics under social suggestions were discovered in a study. Our mathematical research proves the existence of an algorithmic glass ceiling with all of the characteristics of the metaphorical societal barrier that prevents women and people of color from achieving equitable representation. They show that the algorithmic effect is systematically larger than the glass ceiling caused by the spontaneous evolution of social networks when minority and homophily parameters are fixed. But the above analysis did not consider the temporal analysis in glass ceiling effect. To our knowledge, an analysis of temporal effect and the effect of glass ceiling effect in graph has yet to be conducted.

2.2. Temporal social network

Time factor and the characteristics of social media users' social influence pattern in terms of temporal effect are our primary concerns in this research. In this chapter, we are mainly focused on the time factor in the social network.

2.2.1. The basic concepts of social networking

The authors [54] mentioned that, Given a social network, which can be abstractly represented by a directed graph *G = (V,E)*. In directed graph G, V is the set of nodes in the graph, and each node $V_i \in V$ represents a user in the social network. E is the set of directed edges in the graph, and each directed edge $e_ij \in E$ represents the node V_i to V_j. The directed edges of nodes represent the relationships between users in social networks. In this study, we assume that the information in the social network can spread from one user to another user in a directional direction along the edge. As the information spreads between users, the nodes in the social network are affected by the information. Will be in one of two states: The activated state or inactive state is in an inactive state before the user is affected by information transmission. Therefore, all nodes are located in the inactive state at the initial moment. The node initially affected by information, namely the initial activated node, is selected by a human. We call the set of initially activated nodes the seed set. From the seed set, information is transmitted in a certain way. The information changes from inactive to active when an inactive node is affected by the information. There are two characteristics of this process. One is that only active nodes can spread information to influence inactive nodes, and vice versa is not feasible; Second, changing the inactive state to the active state is irreversible. When the node is activated, the state will always remain.

Based on the above assumptions, the influence propagation function can be defined formally to maximize the influence force propagation $\sigma(S)$. It means the expected value of all nodes in the activated state in the social network after the seed set S carries out the influence propagation process. It means that after seed set S carries out the influence force propagation process, the ultimate problem of maximizing the prospective value of the influence force propagation of all nodes in the activated state in the social network is the constraint K of the node in a given seed set. It aims to select the optimal seed set $S \subset V$ so that the value of the influence force propagation function $\sigma(S)$ can be maximized.

2.2.2. The influence maximization problem

[48] mentioned that the influence maximization problem was first stated as an algorithmic challenge, with a probabilistic solution proposed. In 2003, [26] modeled Influence Maximization (IM) as an algorithmic issue for the first time. This problem investigates a social network represented as a graph G = (V, E), where V denotes the number of nodes in G (i.e., users) and E means the number of (directed/undirected) edges in G (i.e., social links between users). The purpose of the IM problem is

to locate a k-sized collection of users in graph G who have the most influence. They also provided a greedy approach that ensures the influence degree stays within (1- 1/e) of the optimal influence degree. On the other hand, the simulation-based greedy approach is computationally expensive and does not scale well in extensive social networks [30, 11].

Following that, other researchers focused on more scalable approaches. Leskovec and others [30] mentioned that the cost-effective lazy forward (CELF) algorithm is an optimal greedy algorithm that reduces computing the influence spread.

The users' information dissemination process determines the influence of any seed set. Viral marketing is an example of information diffusion. A corporation may aim to spread the acceptance of a new product from a few early adopters through social linkages amongst consumers. We formally construct the diffusion model and the influence extended to quantify information diffusion.

In their article [26], they established the two most widely studied models of interaction force propagation in social networks: The independent cascade model and the linear threshold model are referred to as IC model and LT model, respectively. In their study, the target function of the IM problem was given as the expected value of the number of users of the impact of the influence propagation of the seed set, which was later called the influence propagation function $\sigma(S)$. They prove that the IM problem is NP hard under the IC, LT model, and the propagation functions of the influence forces $\sigma(S)$ has the property of monotone submodules. In this paper [26], much actual work has been done. The monotone submodular properties of object functions under some propagation models have been proved, pointing out future research direction.

Although the simple greedy algorithm scholars propose in the article [26] has an approximate guarantee, monte Carlo simulation is used to approximate the incremental value of the transmission function of influence after the addition of candidate elements in each iteration of the algorithm process. As a result, the time complexity of the algorithm is too significant to be applied to real large-scale problems.

Intuitively, the diffusion process would strongly influence the influence function. A considerable body of work has been published in recent years that creates diffusion models to model the diffusion process and compute the influence spread. In the following part, we'll go through some commonly utilized models.

2.2.3. The social influence diffusion model

[33] mentioned that There has recently been a lot of research on developing diffusion models in data mining, databases, networks, and epidemiology. This section examines the models that are typically used for IM, as the focus of this survey is on algorithmic elements of IM.

The reviewed diffusion models are initially presented in a generic diffusion framework. The framework assigns each user $u \in V$ an inactive or active status. The following diffusion process among users is then considered based on G's social graph. Initially, it considers the status of a selected group of users, known as seed set $S \subseteq V$, to be active, while the status of other users in V is inactive.

The diffusion process is then taken into account, with the seed users in S being able to "encourage" their neighbors to become active, newly activated users being able to start their neighbors further, and so on.

The diffusion process ends when there are no new users can be activated. The framework, in particular, the above-mentioned "activation" is modeled as a stochastic process. The stochastic process with the influence spread $\sigma(S)$ being defined as the predicted number of users with active status when the diffusion process ends.

Different models employ various strategies to represent how a user transitions from inactive to active condition, influenced by its surroundings. The Independent Cascade (IC) model, Linear Threshold (LT) model and Time-Aware model are the only three representative models that are often employed in the IM problem.

2.3. State of arts IM algorithms

In this thesis, we are compared different diffusion influenced model, particularly interested in independent cascade model and time-aware model. Since we are going to use these model in our research, understand more will give us sound background and related work in this field.

2.3.1. The independent cascade (IC) Model

The Independent Cascade (IC) diffusion model is a well-known and well-studied diffusion model [20]. It assumes that a user v in social network is activated, and the user v also independently by each of its incoming neighbors by assigning an effect probability $P_{u,v}$ to each edge e = (u, v). Given a seed set S at the beginning of the diffusion process, for example, at time step 0, a diffusion sample of the IC model develops in discrete steps depending on influence probabilities. With probability $P_{u,v}$, each active user u in step t will activate each of its outgoing neighbors v who are passive in step t_1. The activation process is analogous to flipping a coin with a head probability of $P_{u,v}$: if the result is head, v is activated; otherwise, v remains inactive. It's important to remember that u just have one chance to activate its outgoing neighbors. After that, u remains active, and the activation is turned off. When no more nodes can be triggered, the diffusion instance ends. The expected number of activated nodes when S is the initial active node-set, and the above stochastic activation process is implemented the influence spread of seed set S under the IC model.

2.3.2. The linear threshold (LT) Model

Granovetter and Schelling [13, 22] introduced the Linear Threshold (LT) diffusion model in 1978, and it is considered a pioneering model. The core principle behind LT is that if a "sufficient" number of inbound neighbors are active, a user's status can be changed from inactive to active. Each edge e = $(u, v) \in E$ in the LT model is formally connected with a weight $b_{u,v}$. Let N_I(v) be the set of user v's incoming neighbors, and it must satisfy $\sum_{u \in N_I(v)} b_u, v \leq$ 1. Furthermore, each user v is linked to a threshold θ_v. The LT model samples the value of θ_v of each user v uniformly at random from [0, 1] while considering an instance of the diffusion process. Then it moves on to the next phase in a logical order. It sets the status of active users in S to active and inactive users to inactive in step 0. Then it iteratively updates each user's status: All active users in step t1 remain active in step t2, and any user v who was inactive in step t 1 becomes active if the total weight of its active neighbors in $N_I(v)$ is more significant than θ_v. When there are no more users to activate, the diffusion instance ends. The impact spread of seed set S under the LT model, i.e., $\sigma(S)$, is the expected number of activated nodes when S is initially activated, given many instances of the diffusion processes.

2.3.3. The time-aware diffusion models

IC and LT are time-insensitive diffusion models in which the diffusion stops only when no more nodes can be activated. However, propagation initiatives are frequently time-sensitive, and they must optimize the spread of influence while adhering to a strict timeline. To meet this demand, time-aware models have been proposed, and existing research can be divided into two categories: 1) discrete-time models, in which diffusion occurs in discrete steps, and 2) continuous-time models, in which the process of one user influencing another (i.e., diffusion) occurs over time.

The discrete-time models [10, 28, 35] build on IC by representing the diffusion process from one node to the next as a discrete random variable with distinct time steps. Nonetheless, because diffusion occurs in discrete phases, these models are quite similar to IC and LT.

The authors [32] focus on three aspect problems, algorithms, and application in this work. Social graph and diffusion influence model, algorithm: Introduced simulation-Based, Proxy-Based, and sketch-Based algorithm Application. They also mentioned Time-Aware IM, as time factor is an exciting parameter in our research. The research proposes time constraints on the diffusion process. However, they are missing the detail analysis for each algorithm, and they are also not considering the disparity seeding part, which is critical for us research. [55] mentioned that introducing the graph attention networks (GATs), using the experiment to compare the different algorithms and prove GAT has the best performance. However, this paper presents a graph model based on self-attention and did not focus on the temporal aspect. The difference between GAT and GCN is that GAT is introducing one more weight matrix. But this paper did not mention too much detail about the introducing matrix.

[17] mentioned, rather than directly estimating each edge's propagation probability, we try to learn the representation of each node so that the social impact is reflected through nodes' presentation in potentially low-dimensional Spaces. However, due to propagation observations' sparsity, these methods

cannot effectively estimate the influence parameters of all edges, especially for propagation edges that are not sufficiently observed.

Based on [17] mentioned, because the network changed dynamically, neighbor nodes are not increased simultaneously. The connection is not achieved in a short time. Increase if only a snapshot observation network (the snapshot), can only be observed a system of the period of accumulation, and ignore the node increases the process when and how. Therefore, it established the establishment of the network edge through the sequence of events gradually. The sequential network (temporal network) modeling network neighbor nodes and establishing a connection can better reflect the dynamic change of the network. But, the existing network embedding methods have mostly focused on static networks; that is, the network structure remains unchanged, and the number of nodes and edges will not change. As a result, these approaches usually assume that a node's neighbors are disordered; that is, they typically ignore the formation process of link.

Based on [47] mentioned, the authors focus on the prediction of user-level social influence. Their goal is to predict the user's behavior state given a nearby neighbor's behavior state and her local structure information. They inspired by the recent success of neural networks in representation learning, the authors have designed an end-to-end approach to discover hidden and predictive signals in social influences automatically.But, many methods are primarily designed to predict global or aggregate patterns of social impact, such as levels over time. However, in many online applications such as advertising and recommendations, it is critical to effectively predict each individual's social impact, i.e., user-level social impact prediction.

However, based on [33] mentioned, the goal of these studies is to understand temporal influence behavior from observation data, and no IM technique has been presented based on these models to our knowledge. Because this thesis focuses on the algorithmic aspects of IM, which is the most extensively used time-aware model in IM algorithmic research, we will use data analysis to determine the performance of the IM time-awareness algorithm.

2.4. Time aware IM application

Context-aware IM issues have become more prevalent in recent years. Context-aware IM problems are an extension of the classic IM problem in which the context is taken into account. Topical, temporal, and spatial information are more considered in the IM problems. In this section, we are more focus on how the time-aware features are integrated into the traditional IM problem for supporting novel applications. These are criteria for building new algorithms that will push influence maximization research to new heights.

2.4.1. Influence maximization with time constraint

Wang et al. [57] also attempted to investigate the underlying community structure of social networks. To increase the algorithm's scalability, Liu et al. [35] proposed a greedy algorithm based on the concept of the impact spread path. The study of maximizing influence is always based on models, of which the IC, LT, and voter models are three notable models. Even though other modifications of those models have lately been presented, none of them consider the critical aspects of actual influence diffusion such as time restrictions, total budget, repeated influence, and users' online routines.

The article [10] put forward the time-awareness IM problem and the IC model and LT model with encounter events, referred to as the IC-M model and the LT-M model. For the IM problem, the essential bottleneck lies in the calculation of the propagation function of the influence of the node set. Their study introduced two temporal factors, propagation delay, and time-constraint, for the propagation process of the influence and proved the related properties of two new propagation models based on the above properties. To maximize the propagation of the influence force with time constraints, they proposed two initiation algorithms named MIA-M and MIA-C algorithm, respectively, under the IC-M model. The LDAG-M heuristic algorithm is presented under the LT-M model. Although the above algorithm is fast and efficient, it lacks rigorous theoretical analysis, and the approximate properties of the results are not clear, the dynamic distribution of seeds is also not considered.

2.4.2. Seeding strategy with time constraint

[39] attempted to find the best set of initially activated seed nodes to maximize the influence spread in networks. They propose a temporal social network that evaluates sequential seeding based on time windows and a seed selection method, which is sequential seeding. They Explore the capabilities of a sequential seeding technique in temporal networks.

2.4.3. Influence maximization with dynamic graph constraint

The IM methods studied so far are essentially static: they presume that G and the propagation probability p_e for any e $\in$ E are fixed given a social network G = (V, E). However, real-world social networks change over time, as new friendships form, affecting the impact graph. We shall introduce critical research efforts for dynamic IM in the subsequent sections of this chapter, which incrementally process changes in the social graph.

Aggarwal et al. [1] provide efficient algorithms for finding a seed set S at time t such that its effect at time t + h is maximized, given a graph G and its evolution across time interval [t, t+h]. According to Zhuang et al. [64], changes in the graph may only be identified by probing a small number of nodes on a regular basis.

They create efficient methods for two issues at each time t based on this assumption: (1) probing a set of nodes to create a subgraph G_t where the influence diffusion on the underlying graph G_t at time t can be best observed; (2) selecting a seed set S on the observed subgraph G t to maximize the effect of S on the underlying graph G_t (using DEGDIS [11]). The models employed in [1] and [64] are merely proxies and do not correspond to current diffusion models such as IC/LT or their expansions.

The authors [42] present the first fully dynamic method for IM in graph development under IC. Instead of looking at snapshot graphs at discrete time intervals, their approach might provide the seed set for IM in real-time, even if nodes or edges change. Using RRSKETCH to create an index based on the initial graph begins. After that, two fundamental processes, EXPAND and SHRINK, are taught to add and delete nodes from sketches using re-sampling. When it receives a change in node/edge, it will update the affected sketches by doing either EXPAND or SHRINK. The primary principle of sketch maintenance is to ensure that sampling any node or edge is always evenly at random. Following the sketch maintenance, it will recalculate the sample size and, if necessary, generate new sketches or delete current ones. Finally, on dynamically maintained sketches, the seed set selection executes a maximum k-coverage, which is the same as Phase 2 of static RRSKETCH based approaches like TIM. Although the technique was created for the IC model, the concept might be applied to other models such as LT and TR.

3
Temporal Analysis

3.1. Chapter overview

The first phase consisted of a thorough literature review about currently used theories and social influence research methods or fairness seeding algorithms. The next question we need to think about in this chapter is whether the user's behavior changes over time and how the user's behavior changes over time. If we can better understand the relationship between user influence and time series, it will be better to help us to improve our algorithm in information dissemination, furthermore, help us design a better fairness seeding algorithm for social media. Therefore, the next logical step is to use the data acquired from this paper [50], Which mainly focus on Instagram dataset, a detailed description will be given in the next section, and to implement the analysis method in terms of temporal aspect to establish the fundamental knowledge for women's social influence power.

We analyze the dynamic characterization of female and male interaction patterns in this section, answering whether there is a gender gap in users' social influence and strength in the data set when intensity and degree index are taken into account. The social influence analysis looks into how posts from users of different genders are received. And we also interested in whether there is evidence of a temporal glass ceiling effect, i.e., males are over-represented in higher percentiles. The study focuses on how users respond to other people's posts, for example, how users support each other and who supports who.

First, we created a set of questions to discover and discuss possible reasons for different genders' influence. The social influence on a social media platform for each gender had to be made and answered by analysis. Social influence [56] refers that when an individual changes his or her behavior under the influence of others. Social influence's strength depends on the following factors: social network distance, time effect individual characteristics, etc. We mainly focus on the temporal effect, specifically to study the social influence under the temporal effect, the changing trend of male and female.
[45] mentioned that the selection of evaluation metrics is very important to extract a set of evaluation metrics to accurately characterize the characteristics of each user, as these evaluation metrics are helpful to quantify the social influence of each user and to easily find the most influential top-k attributes. Therefore we promote three questions to explore if the homophily, glass ceiling effect and the rich get richer phenomenon will change or remain same in term of temporal effect. In this section, some characteristics of the dataset are introduced, including the statistical information,to describe some findings of the data and the relevance to our research questions.

3.2. Dataset

The experimental data to be got from the Instagram dataset, which collected the data include social platform users' reactions, such as likes and comments sent by target users on each other's posts in 4 years by authors. [50].

Collecting method. The dataset [50] was obtained using the Instagram API by first collecting a collection of users, beginning with Instagram's founder, Kevin Systrom, and then extracting the list of followers recursively. Following that, The authors gathered data on the posters' perspectives and how they received likes and feedback from other users for four years. The user ID, gender (derived from user names), responding form (likes or comments), and timestamps are all recorded for each interaction. The observation period we sample in this paper spans from Jan 2015 to 2018, starting right from the first week of the dataset began recorded.

Data characteristics.the dataset contains about 8 million unique users and about 38 million interactions between them, based on both likes and comments. We are using two main index in this paper for our metric, indegree and intensity. Intensity, which is the the frequency of interaction among connected users, and indegree, which is the frequency of interaction among connected users without multiple interactions with the same user from same source user.

Data filter. In order to avoid data sparsity problem, we decided to filter out some inactive weeks, to make sure there have enough data to use. Roughly 6 weeks to be filter out after our data filtering.

Drawbacks. Because there are a large number of female users on social platforms, a large number of female users are inevitably gathered in each influence classification. Therefore gender bias is a problem for this paper.

3.3. Basic trend of temporal effect

Table 3.1: Level of difference Indegree female groups.

Users	Minimal indegree	Total # users	% of female users	Average indegree
All users	1	4739715	58.97	7.07
Top 10%	10	491297	61.09	38.67
Top 5%	27	241476	60.56	56.18
Top 1%	67	48652	57.13	119.74
Top 0.1 %	184	4764	47.42	297.84

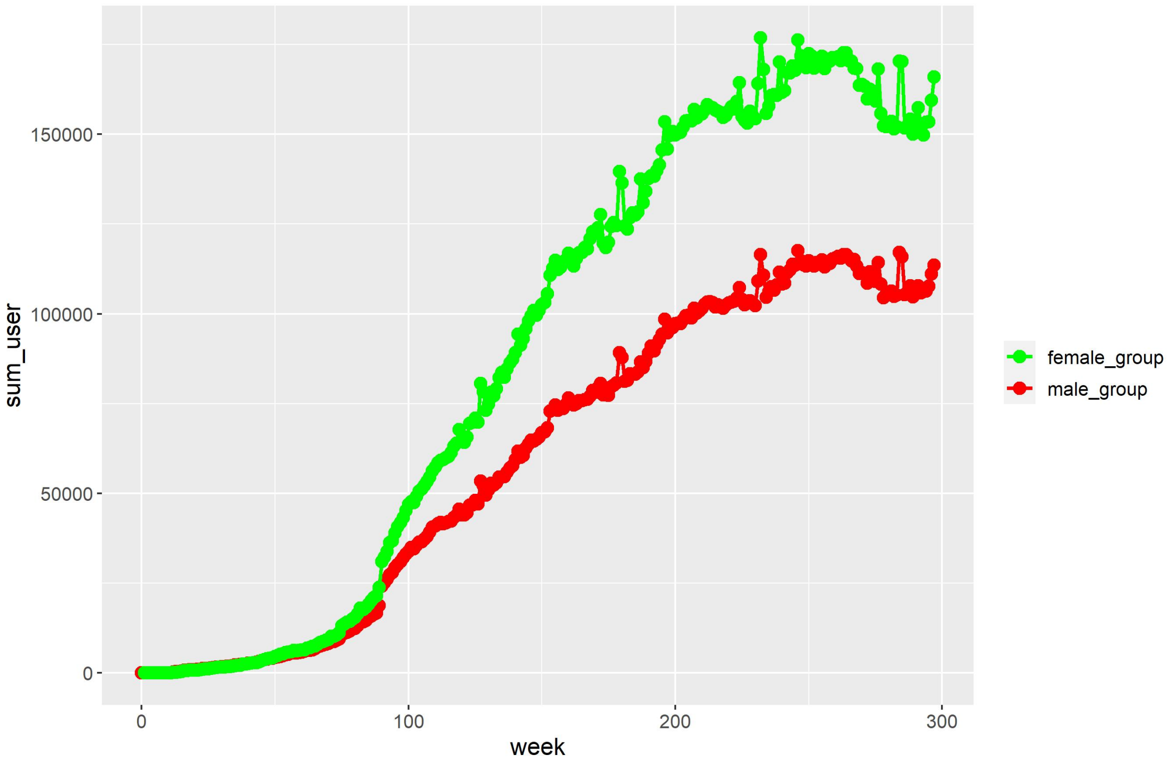


Figure 3.1: Number of active users per week.

There has been a lot of research on the ceiling effect in social platforms, but no too much researcher focus on the glass ceiling effect in terms of temporal effect. [34] mentioned, Most research, however, presumes that the network is static and neglects the role of time in information transmission. Almost all social networks are complex and change over time. When a social network shifts from one time spot to the next, user relationships can be formed or dissolved. When solving the impact maximization problem, in this case, taking into account, the temporal factor will adequately represent the actual structure of social networks. Temporal effect also is the main research direction in our study, therefore, this section focuses on the study of temporal effect and firstly explores whether ceiling effect still exists in terms of temporal effect.

Analysis method. The quantitative analysis method being used to study the distribution of users and average interaction indegree for them. In our temporal effect study, we focused on the top 10% of users, because of [52] mentioned that Females face glass ceiling effects in the social network when measured by PageRank centrality, but they achieve higher visibility for comments. Because the purpose of our research is to improve the algorithm by selecting seeds, users with high ranking have a relatively high probability of being selected as seeds. Meanwhile, users with high ranking also face ceiling effect, so users with high ranking are naturally selected as research objects. The reason for choosing Indegree is that after a large number of data tests, we find that this index is the most representative in reflecting the glass ceiling effect, which can fully reflect the ceiling effect of women in social networks. Lastly, The number of weekly active users is one of the things that we're interested in because it's a good indicator of how engaged and active a social network user is.

The data table 3.1, which shows different levels of users, different levels of influence and different numbers of users. We can see that the percentage of female user decreases as the user level increases, This phenomenon clearly fits the ceiling effect in the social media platform. For the intensity side, we also have done lots of testing, and found the similar trend as well.

The data graph of screened users appearing after week 100 indicates a upward trend of total number of users, suggesting that the overall trend of users is keeping joining the platform and the new users still active after they enter the social platform for a while, as can be seen from the figure 3.1. And after week 200, the trend getting flat, showing that while new users are coming in, the activity of existing users is slowly declining. In addition, it can be clearly seen from this figure that the proportion of male and female users is different over time, which shows that the ceiling effect still exists, which is a powerful proof of the weakness of women in the social field.

3.4. Does the male with higher rank increase the female rank through direct interaction?

Table 3.2: Percentage of user activeness duration: Intensity vs. Indegree.

	Intensity		Indegree	
	Male	Female	Male	Female
Top 0.1% users in gender & in total users	0.415 (± 0.162)	0.528 (± 0.323)	0.530 (± 0.197)	0.642 (± 0.406)
Top 1% users average duration & std	0.373 (± 0.143)	0.461 (± 0.284)	0.435 (± 0.163)	0.520 (± 0.328)
Top 5% users average duration & std	0.35 (± 0.148)	0.403 (± 0.247)	0.368 (± 0.141)	0.438 (± 0.272)
Top 10% users average duration & std	0.333 (± 0.143)	0.382 (± 0.231)	0.345 (± 0.136)	0.411 (± 0.252)
All users average duration & std	0.2959 (± 0.1386)	0.339 (± 0.1973)	0.312 (± 0.145)	0.356 (± 0.208)

Homophily or the propensity of people who identify with each other is one of the most well-established findings in social science. Despite how far research has gone in this domain, we know little about how personal characteristics influence differences in homophily ability [9]. This section focuses on the social interaction of the homophily groups between males and females. By using the method of quantitative analysis, we aim at investigating the interactions among high-end users, such as how female users are more likely to interact with high-end male users. The users we have chosen are divided into four groups: top 10%, 5%, 1% and 0.1% users. For our target audience, We chose top women with a robust social impact. We wanted to see whether their influence comes from top level communities with which they are associated or from the non-top communities which also affect their choices. As a result, in order to solve this issue, we try to understand the characteristics of high-ranking user groups through analysing their social interaction pattern. The group that the target user interacts with is an important factor for this research.

The column entries of table 3.2 represent The percentage of top 10 percent men and women by having the total interactive population divided by the gender total population, which means total female users or male users who interacted with our target group in the same week. The information we gathered shows us how our target female users communicate with the highest-ranking users. For example, To know how many top level users interact with our target users, and also to explore whether high-rank females prefer interacting with high-rank females or males.

The rows of table 3.2 represent the percentages of users activeness duration: intensity and indegree, which correspond to different levels of top users. Top users that are ranked from top 0.1% to top 10% based on their social influence strength.

According to table 3.2, the interaction that top ranked females receive from female users is relatively higher than the one they receive from male users. Because females are found to be 1.5 times more than males and take a larger proportion of the total population in this experiment, the interaction of female users is more than male users.

Given the high proportion of women in the statistics, we would expect females to outnumber male when it comes to the distribution of male and female. Even though the number of females takes a big chunk of data, the percentage of interaction they receive is very close to that male receive; they are taking the lead but with a slight difference. Because the number of interactions among female users is nearly 20 percent higher than that of male users in terms of intensity, it is only about 10 percent higher in the number of high-level influence interactions.

We can see that as high-ranking females fall further and further down the rankings, so does the proportion of high-ranking men they interact with. And the same trend was seen in the percentage of interactions with top women. The data shows that the top 10% of both male and female users interact with the top 0.1% ranking target group, accounting for approximately 50 percent of the total number of interactive users. On other hand, The data shows that the top 10% of both male and female users interact with all users, only accounting for around 30 percent of the total number of interactive users. So the top level women didn't show homophily with all the women, but they showed it with women who are also top users.The results' analysis showed that the percentage of interactions with the top level is very similar between male and female interactors. This means the homophily effect is very strong at the level of top level users. With the drop of users' rankings from 0.1% to 10%, we notice that top ranking male users' percentage of interaction with top level users drop as well.

After our analysis from our result, we found that the percentage of interactions with top level is very similar between male and female interactors. That means homophily exist in the top-top interaction pattern.

We also can see that the percentage of top men and women interacting with leading men and women decreased as the influence of each group getting less.

Anova test has been implemented in this analysis for validation. The p-value result that has been found is lower than 0.05, which proves that the alternative hypothesis is correct. Namely, there are significant differences between different user groups and different categories of users that have a significant impact on the proportion of high-ranking users in the interactive group. Therefore, we can accept my samples and give reasonable evidence to support my conclusion and analysis.

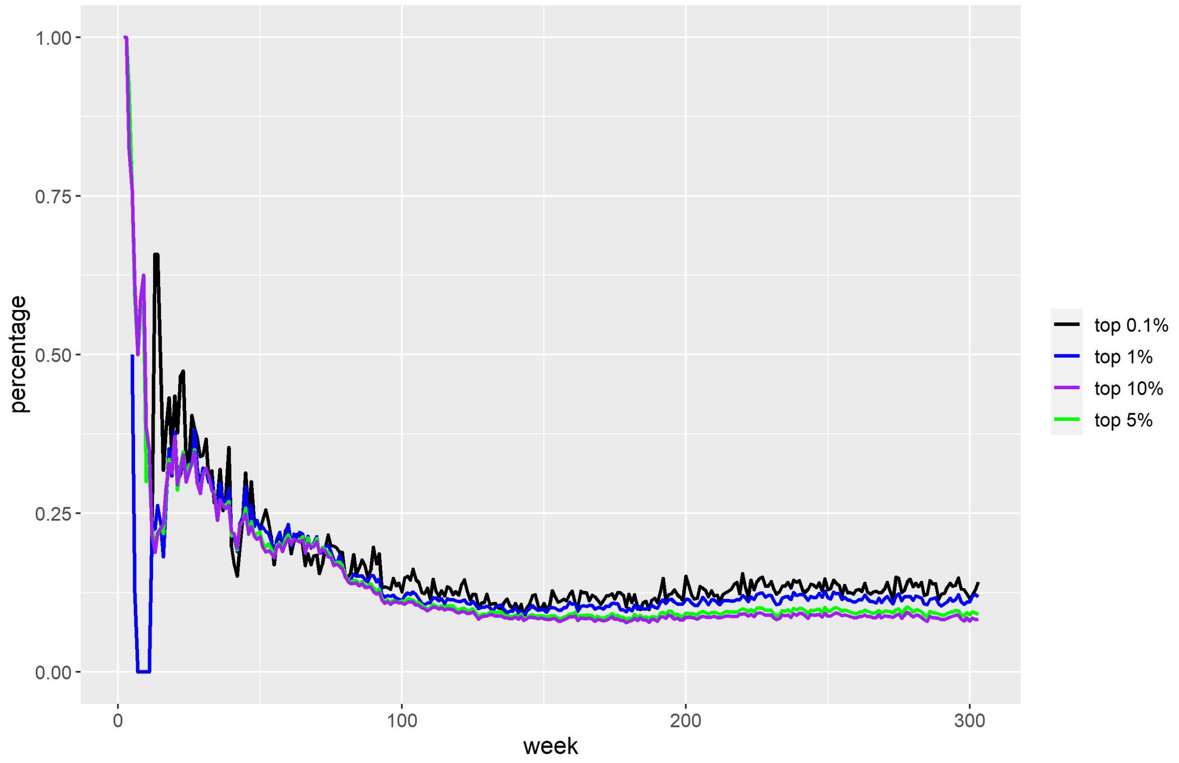


Figure 3.2: Percentage of Male intensity duration over time.

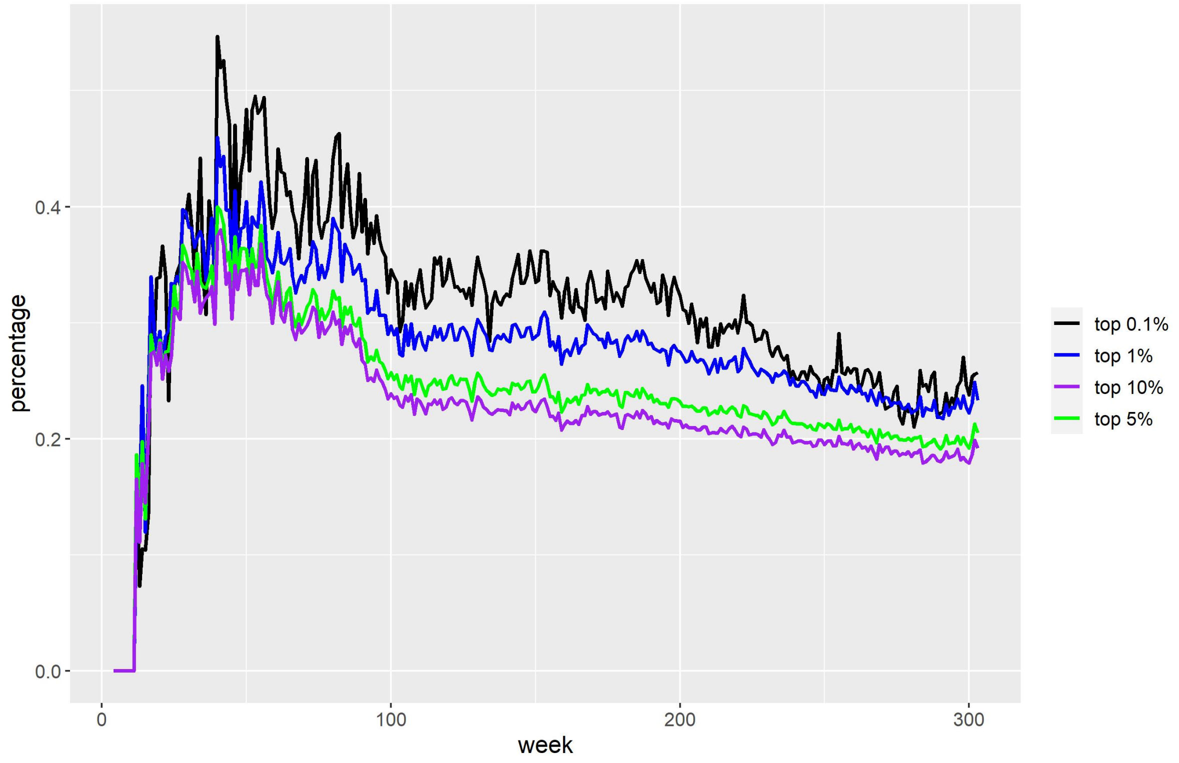


Figure 3.3: Percentage of Female intensity duration over time.

Another interesting question is whether homophily effect changes over time. We analyzed the percentage of highly ranked users in different user groups over time to explore this question. We try to measure the same entries from the table 3.2, and the same target user group as well, in order to realize the influence of time on the research problem as far as possible without changing other observation variables.

From the figure 3.2, we can see that The proportion of highly ranked males in the total number of interactions has a high ratio; the higher ranked target user groups have a high ratio while low-ranked users have a low ratio. However, the trend hasn't changed over time.

From the figure 3.3, we can see that The proportion of highly ranked females in the total number of interactions has a similar trend with the figure 3.3. The higher the ranking of the user group, the higher the interaction percentage and vice versa.

Based on figure 3.2 and 3.3, we can see that Over time differences in the percentage of highly ranked users in the whole interactive users among different user groups remained unchanged. High-ranking target users still have a relatively high percentage of high-ranking interactive users. This means homophily effects are maintained at the same social influence level.

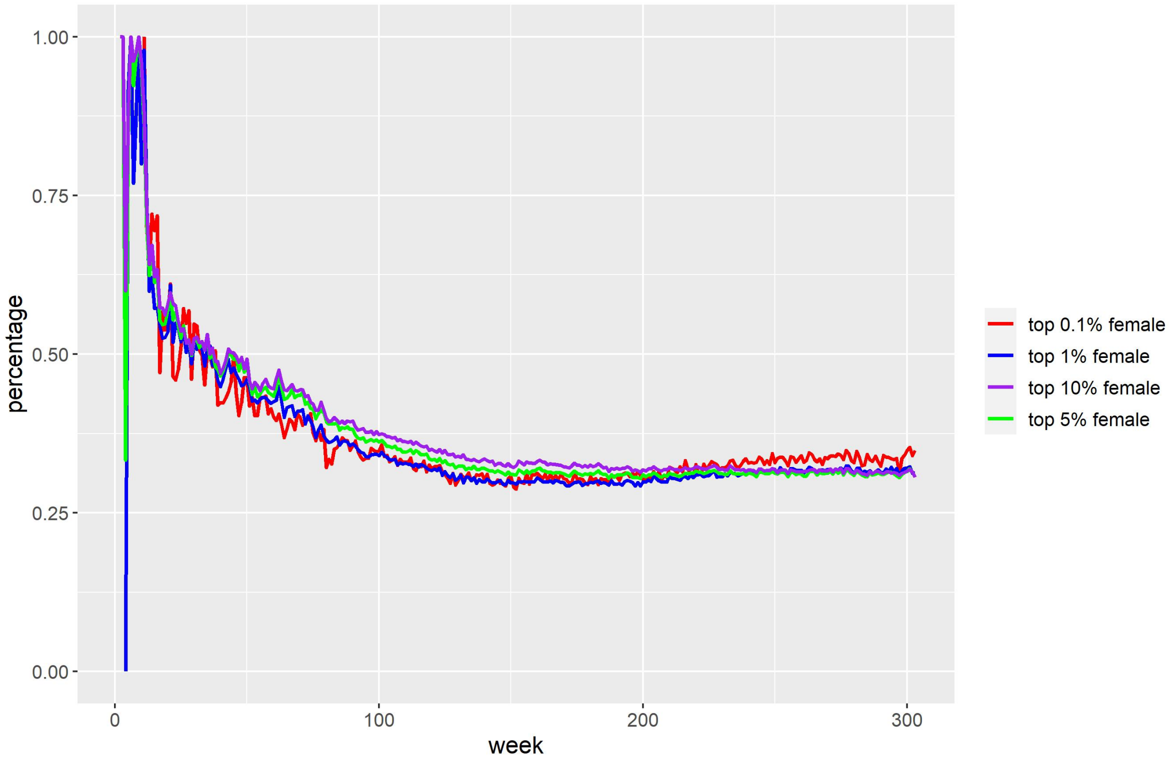


Figure 3.4: Gender ratio in terms of intensity.

Recent study by Twenge and Martin [24] investigated gender differences in the use of social media by examining 13- to 18-year-old adolescents in the U.S. and UK. Results displayed that adolescent girls spent more time on smartphones, social media, texting, general computer use as compared to boys, however, no further investigation was made about how does the gender ratio changes with the time change for the same group of users. Therefore, how does the gender ratio of the interactive users change over time, that's what we're interested in.

From figure 3.4, the horizontal axis shows the time series, from the first week to the last week. And the y axis is the gender ratio is the ratio for these users who interacting with target females per week. The gender ratio equal to male who interacting with target females divide both males + females who interacting with target females. With the time increase, the level of % of gender ratio(males/total population per week) is decreasing, and we also can see that the 0.1% top level of females has a higher level of % gender ratio after week 200, the reason for this situation maybe the higher level of % users still have strong intensity or indegree after week 200. According to figure 3.4, we can see that the gender ratio reach a relatively stable status and the gender ratio is around 0.4. That means with the time change, the target female more willing to interact with male rather than female. This result also show homophily did not strongly apperent between same genders, Because the interaction between the same gender group was not over 50%.

For intensity and indegree, a clear homophily effect can be observed: homophily effect is not strongly apparent between the same gender, however, it is strongly shown between users with similar ranks.

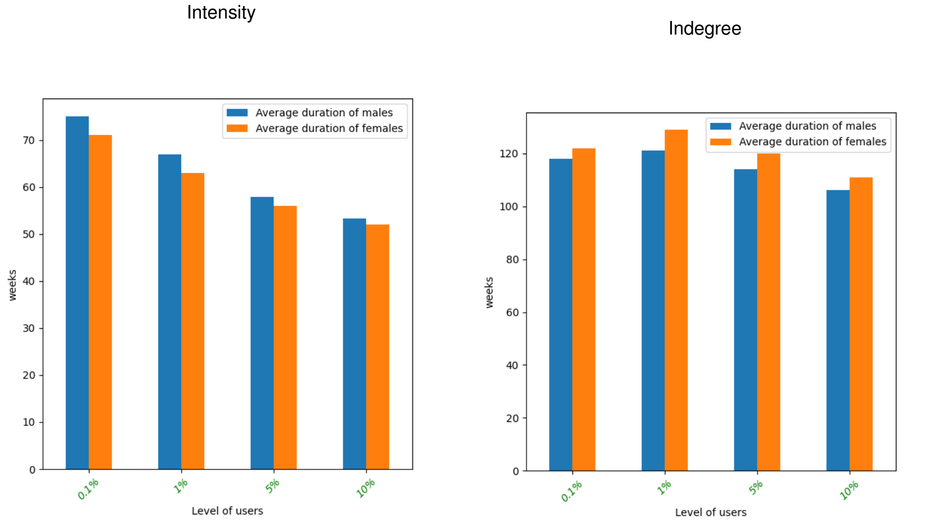
3.5. How long does it take for top users to reach their most influential week?

Table 3.3: User activeness duration: Intensity vs. Indegree.

	Intensity		Indegree	
	Male	Female	Male	Female
Top 0.1% users average duration & std	75 (± 55.68)	71 (± 52.89)	118 (± 68.98)	122 (± 66.02)
Top 1% users average duration & std	67 ± 51.40)	63 (± 49.47)	121 (± 65.52)	129 (± 63.12)
Top 5% users average duration & std	58 (± 48.25)	56 (± 46.93)	114 (± 60.61)	120 (± 57.68)
Top 10% users average duration & std	53 (± 46.73)	52 (± 45.50)	106 (± 58.26)	111 (± 55.81)

In order to maximize social information propagation, the characteristics of the selected seed users are important. We want our chosen seed users with their greatest impact at that moment to maximize our goals. The characteristics of the top users in the social network, such as the time it takes to reach their max social influence, are becoming our interest point. So the main purpose of this section is on when the user's influence is maximized. We focused on the average time it took users to reach their maximum and grouped them by their level of social influence.

From the data table 3.3 we can see that average duration of females are higher than male, which is reasonable and the reason cause to this situation is the influence of male growing faster than female. And there is another trend that we can read from the table 3.3, on the top level, top influence of males have higher social influence than female, that means the higher the impact, the longer it takes to reach their peak impact. And from these data we can see that with the top percentage of people increasing, the average duration of reaching top level getting lower, it is also confirmed by the above idea. We also can see that from indegree measure,the average duration of females takes longer than male. Given our previous data, women are generally with less social influence than men, that means the slope of women reach their peak time are lower than men.



(a) Activeness duration for top users to reach maximum influence week

(b) Activeness duration for top users to reach maximum influence week

Figure 3.5: The Activeness duration for top users to reach their maximum influence week.

Furthermore, the figure 3.5 reveals another trend: users at the top of the list have a higher degree of dispersion, while users at the bottom of the list have a lower degree of dispersion, implying that the prediction of the time when a highly ranked user reaches the peak of influence may have higher volatility. This reveals that using a complex network to pick seed users could be more accurate than using a range interval, since the time it takes for high-ranking users to reach their peak of control varies greatly and is relatively difficult to predict.

> The intensity and indegree of the highest influence to be achieved can be clearly observed, the more influential the users are, the longer it takes for them to reach their peak of influence.

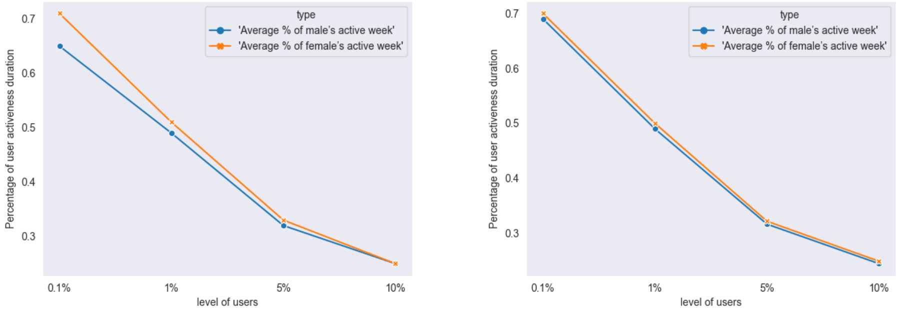
3.6. How can disparity of users' activeness and gender yield different active durations?

Table 3.4: Percentage of user activeness duration: Intensity vs. Indegree.

	Intensity		Indegree	
	Male	Female	Male	Female
Top 0.1% users average duration & std	0.69 (± 0.121)	0.70 (± 0.121)	0.65 (± 0.197)	0.71 (± 0.162)
Top 1% users average duration & std	0.49 ± 0.156)	0.50 (± 0.155)	0.49 (± 0.173)	0.51 (± 0.165)
Top 5% users average duration & std	0.3166 (± 0.16)	0.3219 (± 0.16)	0.32 (± 0.164)	0.33 (± 0.163)
Top 10% users average duration & std	0.2447 (± 0.153)	0.2491 (± 0.153)	0.25 (± 0.155)	0.25 (± 0.155)

For the change of user influence in terms of the temporal effect, user activity is also a focus of our attention. Looking at the activity levels of different ranking user groups can also give us some insight, such as whether the more active users are, the more influential they are. Based on [4] mentioned, they made an arguments about the rich get richer mechanism, this mechanism predicts that people will want to link more frequently to people who already have many links, either to benefit from their social wealth or because they are more visible in the network, in our setting where the degree of the vertex captures its level of social wealth. Given the specifics of social networks, whether the more active users are, the more they interact with other users, and so they have the Matthew effect in the social influence level. To explore this question, we mainly focus on the average activeness factor for each target user group. The average activeness factor means that the number of users being active in the specific week divided by the whole users in our target user group. And then redo the above step for 303 weeks. By computing this factor, we can clearly know that The level of activity of users in different groups.

From the table 3.4 we can see that as the ranking of user groups drops, so does the level of user activity. With the percentage of people increasing, the average % of both female male's active weeks are decreasing, So we can come to a conclusion that user ranking is positively correlated with their activeness.

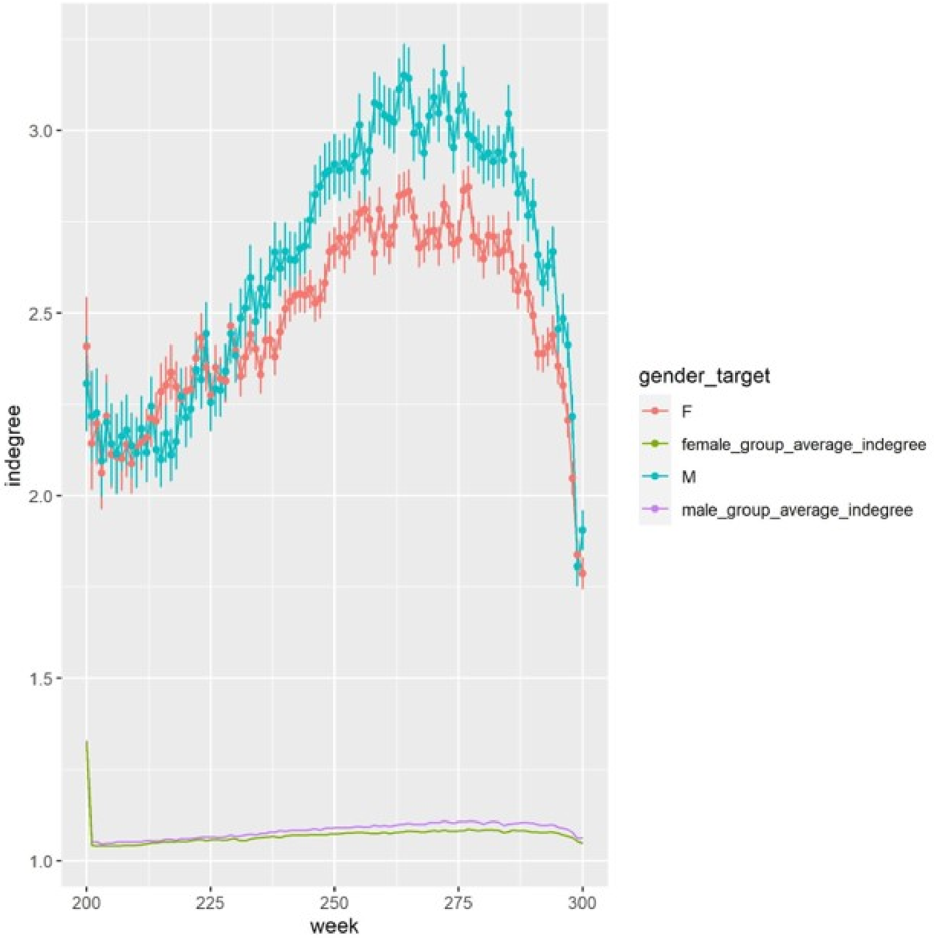


(a) Normalised active duration in intensity

(b) Normalised active duration in indegree

Figure 3.6: Normalised top users' active duration.

According to figuree 3.6, by analyzing the standard deviation of user activity, we can see that the higher the user ranking is, the more stable their active dispersion is, indicating that the higher the user ranking is, the higher the probability of user activity is.



(a)

Figure 3.7: Average indegree for both male and female over time.

After investigation, we found that women are more active than men, but whether the average intensity or indegree of women will also exceed that of men as well? To research this question, We calculate what the average intensity of each week is for different ranked user groups, furthermore to explore the change trend of average indegree for each user group. In this case, after lots of testing and analysing, we found that indegree can represent the average trend clearly among of other graphs. And the user we choose is the New users to join after week 200, the reason for this is our study found that the user's activity level will decrease with the decrease of joining time. If we choose users to join too early, our average intensity results will be inaccurate, which will affect our conclusion. After 200 week is a good time slot because in this period of time, the newly joined users were quite active, and with the increase of the platform's opening time, the number of users joining was also relatively large and stable, which ensured that we had enough data to support our research. Therefore due to there are lots of user being inactive after long time join in, to improve our Accuracy of experimental results, we only select users have occurrences after week 200, look at their behaviors between week 200 and week 300. And we can see from figure 3.7, at both the levels of male and female , men had higher average social intensity than women for intensity and indegree. Men exhibit better average social influence than women judging from the data that our target audience is presenting.

As can be seen from the above data, although the active time of men is relatively shorter than women, the social influence such as intensity within a period of time is higher than women.

Based on our data analysis from figure 3.7 the y bar means the indegree or intensity for each user, the x bar is the number of weeks. The gender group divided by four parts, top percentage of male or female, male is the blue color and female is the red color, and the average indegree or intensity of the entire female and male group. This chart clearly show how the average influence of top level's men and women has increased over time, and although the whole group also showing the similar trend, Men have more influence on average than women, and this has not changed over time. And combine with the previous result, we can see that despite the number of active user getting smaller, but the exist users in top level showing a increase intensity and indegree trend. This means that the average influence of both highly ranked men and women has gradually increased over time. reveals there is the rich get richer mechanism exist in terms of temporal effect.

The average intensity of both males and females results in an increased trend, representing the rich get richer mechanism in terms of temporal effect.

3.7. Double validation of result of temporal effect

In order to make our experimental results more accurate, we are using double validation method to certify our analysis result. The facebook is a good data set to validate our results, because [61] mentioned that Facebook and other social media sites have become significant channels: according to recent studies, members use these platforms not just for social contact but also as a source of political and public affairs knowledge. Facebook is one of the most important websites today, therefore using facebook to vertify if homophily and glass ceilingeffect in terms of temporal effect also exist.

3.7.1. Facebook dataset

We collected from facebook dataset, which is gathered from sample survey of fresh university graduates. We gathered these data based on the following variables, edge_type(received or send likes or comments), actor_id(person ID), action_time(the action created time), post_owner_id(the person who published this post), ownerFbid(the person whom this data was scripted from), gender.

Collection method. We collected the facebook user data from 25 university departments by using facebook API. These users are study at these universities. *Data characteristics*. the dataset contains about 50000 unique users(all of them are student) and about 14 million interactions between them,based on both likes and comments. We are using two main index in this paper for our metric, indegree and intensity, like the Ins dataset we used for the temporal analysis. Based on [52] metioned, Furthermore, this dataset also gathered user information, such as academic standing and hometown, by questionnaires. The period of iterations spans from March 2008 to May 2016, and 97.26% of interactions are after August 2012.

Data filter. We decided to filter out inactive weeks and the week with too few users to make there is enough data to use, the first 100 weeks and the last 100 weeks has been removed from our test.

Drawbacks. Because of there are much more female than male in the Facebook dataset. This phenomenon should be carefully considered when analyzing data such as the ceiling effect if exist in terms of temporal effect. And the data has been gathered for a long time, but the users groups are quite small. It's a better idea to find a more generalize dataset which can contain more general population of Facebook users.

Table 3.5: Facebook distribution of users.

Users	Total # users	% of female users	Number of male	Number of female
All users	42041	98.70%	545	41496
Top 10%	4205	87.90%	510	3695
Top 5%	2102	87.80%	439	1663
Top 1%	420	74.05%	109	311
Top 0.1 %	42	71.43%	12	30

In order to further study the sample data of Facebook, we decided to analyze the distribution and characteristics of Facebook users, to better analyze the proportion of female users, especially the proportion of high-influence users.

We divided users into different levels according to the percentage of user ranking and influence, and calculated the total number of users for other user groups, the proportion of women in the total number of users, and the number of men and their proportion.

After our test, from table 3.5 we are found that the student sample survey data of Facebook presents a severe female bias. That is, the vast majority of users are female. However, as we can see from the figure, the number and proportion of females decrease with user ranking. The ratio of females reduces significantly among users with a high level of ranking users. This also confirms our previous findings that the percentage of women decreases as level of users influence increases.

Table 3.6: Demographics interacting with top females: Intensity vs. Indegree.

	Intensity		Indegree	
	Male	Female	Male	Female
Top 0.1% users in their own & overall users	0.935 (± 0.073)	0.245 (± 0.233)	0.946 (± 0.238)	0.088 (± 0.069)
Top 1% users in their own & overall users	0.215 (± 0.014)	0.246 (± 0.227)	0.219 (± 0.052)	0.092 (± 0.071)
Top 5% users in their own & overall users	0.124 (± 0.007)	0.250 (± 0.231)	0.124 (± 0.028)	0.087 (± 0.067)
Top 10% users in their own & overall users	0.123 (± 0.007)	0.253 (± 0.235)	0.125 (± 0.027)	0.086 (± 0.066)
All users in their own & overall users	0.123 (± 0.006)	0.259 (± 0.243)	0.125 (± 0.027)	0.086 (± 0.077)

In order to make a comparison with does the female rank higher because the male that they interact with has a higher rank this question, We also simulated the same test as before to see if the sample of Facebook users showed the same or similar trends.

As can be seen from the tables 3.6 and 3.7, our test obtained similar results. Namely, the proportion of highly ranked users in the interaction was also high, while the proportion of low-ranked users and highly ranked users decreased in turn. As can be seen from the figure, although women account for an overwhelming proportion of the total number of users, in terms of interaction, female users show a preference for high-impact male interactions. This feature was also a new found in the testing with Facebook users.

When analysing by using the Facebook dataset, the homophily effect can be observed as similar as Ins dataset, and the Facebook dataset is shown top level of female users prefer to interact with top male users.

Table 3.7: User activeness duration to reach their top influence: Intensity vs. Indegree.

	Intensity		Indegree	
	Male	Female	Male	Female
Top 0.1% users average duration & std	111 (± 60)	128 (± 84)	213 (± 62.29)	222 (± 96.80)
Top 1% users average duration & std	108 (± 66)	102 (± 68)	159 (± 97.44)	171.60 (± 95.00)
Top 5% users average duration & std	93 (± 64)	84 (± 66)	127.99 (± 103.77)	122 (± 99.47)
Top 10% users average duration & std	90 (± 66)	61 (± 60)	118.21 (± 104)	80.69 (± 87.88)

It is also essential to use different data to explore the trend when users reach the peak of their influence. Our conclusions can be verified from several various data sources.
The user groups and variables for the test are the same as before, and the result also we get the similar one. From the table 3.7 we can see the higher the level of users ranking, the longer it takes to reach their peak impact.

Table 3.8: Active duration in terms of percentage of normalized active duration.

	Intensity		Indegree	
	Male	Female	Male	Female
Top 0.1% users in gender & in total users	0.9172 (± 0.08)	0.8911 (± 0.09)	0.825 (± 0.108)	0.806 (± 0.109)
Top 1% users average duration & std	0.8442 (± 0.1)	0.7819 (± 0.11)	0.729 (± 0.129)	0.725 (± 0.133)
Top 5% users average duration & std	0.5506 (± 0.27)	0.4121 (± 0.25)	0.506 (± 0.215)	0.448 (± 0.236)
Top 10% users average duration & std	0.5088 (± 0.29)	.2097 (± 0.24)	0.459 (± 0.239)	0.238 (± 0.250)

Whether user activity is related to user influence is also one of the research points that we are very interested in because user activity can be taken into account in our algorithm design, such as setting different seed screening conditions for different activity levels. Therefore by using Facebook to double validate our previous is necessary. The all test settings and the variables that we select as same as before, and the result we get as the following.

From the table 3.8 we can see that, The more active a user is, the more influential the user is, and the longer the user is active. And we also test the average intensity for each influence level of users in terms of time series, we found the similar result as the previous section that we mentioned, high ranking users become more active compare with the lower level of users.

After verifying the experimental results, we have found that the sample data of Facebook still showed a similar trend to the analysis results of Instagram, it can be observed that homophily effect did not strongly apperent between same genders, however it is strongly showing between the similar influence of difference group of people and the rich get richer mechanism also exist in both social media platform in terms of temporal effect.

4 Algorithm

4.1. Chapter overview

This chapter discusses the approaches used to answer the research questions defined in the first chapter of this thesis. First we describe the reason on choosing the specific evaluation metric as our proxy that is aligned with the objective of dissemination of information. Then, disparity seeding framework will be discussed. Lastly three diffusion algorithm are described: original algorithm, temporal equally seeding algorithm and unequally seeding algorithm. The original algorithm as the baseline method, to explore on the temporal aspect, the performance of benchmark algorithm which based on IM model. The temporal equally seeding algorithm as our alternative method, to explore What happens when the strategy of evenly distributing seeds is adopted, while the unequally seeding algorithm is our proposed algorithm, which is more focused towards answer the research questions.

4.2. Evaluation metrics

The proposed disparity seeding algorithm in this thesis needs to be aligned with the research objectives of our project - fairness of gender ratio and maximize the diffusion of information spread, which technically are represented via several evaluation metrics. The fairness of gender ratio is translated into keeping the influenced ratio as close as the target ratio, which eventually lead to balance the gender ratio in the result of total outreach. The maximization of the diffusion process of the information spread is measured by the total outreach.

A successful disparity seeding algorithm is determined by its performance in an online social communication environment. However, There are many model for studying diffusion model, such as independent Cascade(IC) model, linear threshold(LT) model, triggering(TR) model [33]. However these are all time-unaware model, which means the diffusion terminates only when no more node could be activated. Therefore, performing disparity seeding algorithm in a time-unaware social network graph has already been studies. Instead, time-aware diffusion model has relatively large research space. Due to propagation campaigns are often time-critical and need to maximize the influence spread under a time constraint. Considering this reason, we need find a time-aware algorithm which might have good performance in reality situation.

Note that when considering the temporal factor, users information spread ability differently in different months. e.g. The more influential the user, the longer it takes to reach the peak of his influence than the less influential user, which is the conclusion from our chapter 3. Therefore, under the temporal effect, whether the user's propagation performance is the same as that without considering time condition is a problem worth to paying attention.

Practically, we want to find what kind of metrics we can use to measure the user's information spread ability. and observe how difference metrics change under the time-aware condition, and what is the information spread trend looks like.

In the following, we first identity the metrics we used to measure the visibility. Second, we apply time-aware method into the following metrics.

4.2.1. Interaction intensity

For measure the visibility, we want to know which gender is more influential and which gender is more willing to endorse others based on the volume of comments and likes. For analyze visibility, it is also important that to know the number of interactions between users, by reading previous literature [3], we want to first investigate visibility by two type of metrics, degree the number of interaction partners, and intensity, the number of interactions. For measure the endorsement, we can focus on how users support posts from others to display the endorsement activities, we can count the total number of message sent by unique users. Based on the document research [4] and previous data analysis from chapter 3, we have identified the following characteristic for the visibility intensity and endorsement intensity:

1. The number of times each user interacts is counted, This means that interactions between the same users are counted, for example, When user A interacts with user B in n times over a period of time, then both user A's intensity and user B's intensity are n. With this index, we can clearly describe the number of interactions between users.
2. The level of intensity is easily affected by the interaction between a few active users, which could result in user who with high intensity, but the actual number of user interacted is not such high.

4.2.2. Interaction degree

For masrue the visibility and endorsement activities of target users, we also can use interaction degree also called indegree, as our metrics. The interaction degree is the number of interaction partners, the characteristic of indegree is like following based on the previous research [52]:

1. The visibility degree records that when a user receives interaction from other users, if the other unique users are n, the index increases by n for total n unique users interaction. This means that the number of interactions with difference users is not simply determined by the number of interactions, but the number of unique interactive users is also important. For example, if user A interacts with n users, and each user interacts m times, his visibility degree is n rather than m.
2. The endorsement degree showing that when a user send interaction to other users, if the other unique users are n, the index increases by n for total n unique users interaction as same as the previous definition.

4.2.3. Target HI-index

Initially intended for an individual scientist or researcher, the HI-index is an author-level indicator that assesses publications' productivity and citation effect. The HI-index is the maximum value of h for an author/journal that has published at least h papers, each of which has been referenced at least h times [37]. In other words, the HI-index is based on the number of articles in the citation network. As follows, we extend this concept to account for the strength of interactions, which is the HI-index. The target HI-index derives its definition from the HI-index concept, and the characteristic of target HI-index is also discovered by [52], and I am going to summarize as following:

1. The highest number H such that vi has at least H neighbors who interact with vi and any other users in the network at least H times is defined as the HI-index of the user vi. Assuming that N (vi, n) to represent the number of vi's one-hop neighbors who interact with others at least n times. The formula of the HI-index of vi can be: H(v_i) = $\max\limits_{n} \min\limits_{n \subset I^+}$ (N(v_i,n),n)
2. As a result, the HI-index extends beyond a single hop analysis by analyzing all connections between neighbors in a social network, as well as interactions between a post's author and its supporters. As a result, a user's HI-index is determined by their two-hop neighborhood.
3. The definition of target HI-index have been created in paper [52], A user's target HI-index is defined as (v_i) have been defined as (v_i)'s HI-index, x but penalized by the difference between ζ and the female ratio of (v_i)'s from which the HI-index is generated. Let N^F(v_i,n) denote the number of (v_i)'s direct female neighbors interacting with others at least n times. The Target HI-index of (v_i) is formulated as TH(v_i, ζ) = H(v_i) • (1 - ζ(H(v_i), ζ)), where γ(H(v_i),ζ) = |$\frac{N^F(v_i,H(v_i))}{N(v_i,H(v_i))}$ - ζ| is the penalty for the female ratio of v_i's direct neighbors having at least H(v_i) interactions with others not satisfying ζ. A larger difference between ζ. A larger difference between ζ and the female ratio in N(v_i,H(v_i)) results in a greater penalty on H(v_i).

4.2.4. Pagerank

The PageRank centrality metric, created initially to rank web pages by their popularity, which is another frequently used metric to evaluate the influence throughout an entire network. [59] metioned that PageRank (PR) is a Google Search algorithm that ranks web pages in search engine results. It is called after co-founder Larry Page and "web page." Basically PageRank is a way of measuring the importance of website pages. PageRanks is a link analysis technique that gives each element of a hyperlinked group of documents, such as the World Wide Web, a numerical weighting to "measure" its relative relevance within the set. Any collection of entities containing reciprocal quotations and references can be used with the algorithm. Beyond web search, PageRank is a helpful centrality in a variety of network study applications [5, 21]. The characteristic of pagerank is like following:

1. The PageRank transferred from a specific page to its outbound links upon the next iteration is divided equally within outbound links. For example, If page A has link from pages B, C, and D, each link would transmit 0.20 PageRank to A on the next iteration, for a total of 0.60.

$$PR(A) = PR(B) + PR(C) + PR(D). \tag{4.1}$$

2. Both the HI-index and PageRank use the degree and intensity of interactions to determine the visibility level of nodes. There is a distinction. For a user to acquire a high HI-index value, they must have many highly visible neighbors. Because for the Hi-index, if the user want to get high value in the HI-index, their neighbors also need to have high visible. In other words, the HI-index requires to exceed two thresholds, interaction partner quality, and quantity. In PageRank, That means a user can be popular even though their number of neighbors is small if their neighbors are also prominent.

4.3. Disparity seeding

In this section, we'll consider how to combine temporal factor, aforementioned research in chapter 3 with seeding approach and then use this approach to maximize the transmission of information for a specific demographic group. According to the aforementioned study in chapter 3, when selecting seed users, consider that the level differences among seed users have a more substantial impact than the user's gender. Based on our previous research, we can infer that directly selecting seed users without filtering will lead to an imbalanced gender ratio for seed users if we use these seed users to do the influence spread. However, the relative proportion of male users with a high ranking is higher than that of female users. This could lead to the gender ratio imbalance for the influenced users. Therefore the evaluation metrics to be considered to help us rank seeds users fulfill our requirement.

The time factor also is an important factor need to be concern. For example, we can try to make our influenced ratio as close as the target ratio when we need to consider temporal issue. In traditional IM algorithms, each diffusion instance ends only when there are no more nodes to be influenced. For example, it may take O(n) steps in discrete time diffusion models, and the process could take any amount of time in the continuous time models. To establish a temporal constraint on the diffusion process, time-aware influence maximization (TimeIM) is presented. Because the diffusion process can take a long time to stop, this assumption is practically illogical. This approach has more practical meaning than the previous strategy[52] because in real life, the network graph is dynamic rather than static. The solution can be used in commercial and government initiatives that aim to reach a specified percentage of females in the most effective way possible [53] or other situations used to regulate gender ratio. Usually, when companies need to advertise in the market, these entity usually select a fixed number of users to help them achieve their company goals. The number of female users as seed are relative low because of the glass ceiling effect. Based on the previous paper [52] mentioned, and we extend the definition of the problem of influence maximization with a target gender ratio restriction under temporal aspect as follows:

Definition 4.3 (time-aware Disparity Influence Maximization (TADIM)). Given a series social network from G_1 to G_n, where G_i = (V_i,E_i), $i \in n$. And a diffusion model, a seed group size K, a set of seed size K_1 to K_n, $\sum_{i=1}^{n} K_n$ = K, a target gender ratio ζ, and an error margin e, the problem is to select a seed group $S_i \in V_i$ with $|S_1|$ = K_1 ... $|S_n|$ = K_n to increase the spread of influence within the constraints of a female-to-male ratio in the influenced users is ζ, within an error margin e.

Note that [52] metioned that DIM is NP-hard since it is as same as the traditional IM problem, which is proved as NP-hard, where e = $+\infty$. In our case, the TADIM it is also an NP-hard problem.
We introduce a theoretical model that reproduces the difference between different communities in social networks better to understand the underlying reasons for bias in influence maximization. We can investigate the complex relationship between diversity and optimally by mathematically analyzing the conditions in which bias is first created and then captured by algorithms that leverage network centrality while simulating the organic growth of communities of individuals with different features or interests.

4.3.1. Time aware disparity seeding framework

The time aware disparity seeding framework is proposed, selecting important females and males based on a seed number and a goal gender ratio that may differ significantly from the population ratio. We investigate two dissemination networks, namely commenting and liking, and use a simulation-based strategy to efficiently attain the intended ratio for a given network, motivated by the gender as mentioned above variations in interaction patterns. We look at two types of dissemination networks: commenting and like. and based on the simulation-based approach to achieve the targeted ratio for a specific network. Our disparity framework is composed of five phases, illustrated in the following figure:

1. transfer graph to time aware graph (preprocessing)
2. Allocate the size of user seeds for each graph based on predicting graph size and clustering coefficient of each graph
3. Rank users for each graph
4. Deciding seeding ratio
5. Diffusing information

In the phase of transfer graph to time aware graph, we first generate difference graphs based on Different time periods. For example, In this experiment, we selected 24 months as a complete experiment cycle and divided each month into a period, the total lengths of the experiment cycle is T, and the first month of the the time period is T_1, the second month of the time period is T_2 and so on until T_n.

After the first phase, we will generate graphs of different sizes based on different periods and figure out what proportion of seeds belong in the total seeds limit based on the size of the other graphs. The proportion of seeds in each graph is S_i, $S_i \in$ S, the total number of seeds is S. We also introduced a parameter,clustering coefficient, as one of our factors for determining seed proportion in the total seeds. Regarding [58] mentioned, The clustering coefficient is a measurement of how closely nodes in a graph cluster together. Evidence reveals that in most real-world networks, especially social networks, nodes form tightly knit groupings with a relatively high density of links; this likelihood is higher than the average probability of a tie forming randomly between two nodes. Then we generate a compound parameter - the seed selection parameter, which combines the previous graph size factor with the Clustering Coefficient factor, combining the two factors based on a specific proportion for each element. At the third phase, we rank users by identifying influential users according to their ranks with HI-index, PageRank, indegree or intensity.

Then, in the deciding seeding ratio phase, we estimate how much of the available seeds to assign to males and females for each graph, abbreviated as the seeding ratio, depending on their ranks. Our previous investigation revealed a significant disparity between the two, with a higher percentage of very influential guys despite a lower ratio of males in the population. And another fact also need to pay attention to is our data analysis found that the proportion of female users was much more significant than that of men, so female users were more likely to be selected as the seed because the number of female users accounted for a high proportion of the total number of users. A diffusion simulation on a tiny seed set is used to capture the relationship between the target ratio and the seeding ratio.
Lastly, We Run a diffusion simulation based on a temporal graph to estimate the information spread according to the seeding ratios for the two types of Disparity Seeding in the phase of diffusing information.

4.3.2. Defining Stages of time-aware disparity seeding framework

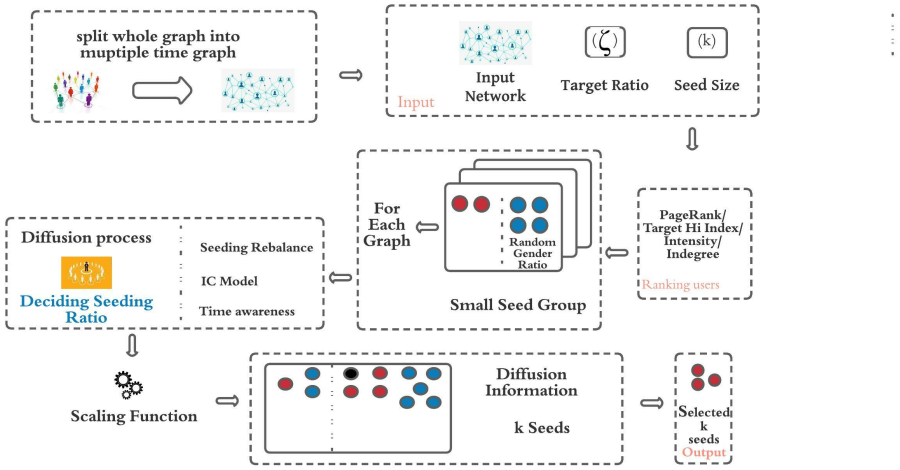


Figure 4.1: The time-awareness disparity framework

i) Transfer graph to time aware graph
The preprocess is similar with the original paper[52], but with more consideration with time factor. The whole process following the steps:

1. Generate the graph based on time period
2. Generate value of users for each index in different graph
3. Generate seeds based on time-awareness disparity seeding framework

ii) Allocate the size of user seeds for each graph based on predicting graph size and clustering coefficient of each graph

Table 4.1: Notation of time-awareness seed selection formula

Notation	Definition
S_i	seed size per graph
G_i	size of graph
W	weight factor
C_i	cluster coefficient per graph
T	total time period

This step is newly created method, we found that if the distribution of seeds is distributed evenly according to the period, it will not conform to the actual user activity and transmission, thus affecting the transmission efficiency of seeds. Therefore we compute graph size and the clustering coefficient and combine these two index together to generate the new index. Then we distribute the seeds based on the new metrics following the formula:

$$S_i = S * (G_i * W + C_i * W) * (1/T) \tag{4.2}$$

iii) Rank users for each graph
The ranking process is as same as the original [52], but with the temporal factor consideration. The previous paper sort users based on the original graph, while we decide to generate the sub graph for each time period (different time period has the same time length), and rank the user based on each sub

graph rather than the original graph. In this way, we can screen out the active users of the week and sort the active users, which will be more efficient when spreading as seeds.

iv) Deciding seeding ratio

Instead of picking 10 numbers from zero to one, we're randomly generating the seed ratio so it's a lot more random and it's a lot more robust than the previous method. It has the advantages of being free of categorization mistake and requiring little prior knowledge of the population other than the frame. Because of its simplicity, data acquired in this method is generally easy to interpret. For these reasons, simple random sampling is best suited to situations in which little information about the population is available and data collection can be done efficiently on randomly distributed items, or where the cost of sampling is low enough that efficiency is less important than simplicity. If these criteria aren't met, stratified or cluster sampling may be a better option [40].

v) Diffusing information

The diffusing information we create two new methods, equal diverse seeding algorithm, unequal diverse seeding algorithm and one basedline method, origin diverse algorithm. The specific detail for diffusion process will be introduced at next section. In this part, the general diffusion process would be discussed. The diffusion algorithm based on IM Model, and it has the characteristics of time awareness, which means the input of graph is not the whole static graph, instead, the graph is segmented in time and is continuous, constituting a dynamically changing graph relative to the previous static graph. The seeding rebanlance method used to help unequal diverse seeding algorithm dynamically adjust the number of seeds and the proportion of gender ratio.

vi) Scaling function

We used linear regression to try to fit the optimal influenced ratio. Specifically, we generated the corresponding seeding ratio according to the selected target ratio, and then we substituted the seeding ratio into the diffusion algorithm and calculated again to obtain the corresponding influenced ratio. The influenced ratio is The relative optimal ratio that we expect.

vi) Generate K seeds

The total outreach of influenced users has been generated after the diffusion process. And the absolute error is also to be obtained after the scaling function and diffusion process. The k seeds would be generated by comparing the different running results and based on our multiple running, and we will select the best K seeds based on our algorithm.

4.4. Algorithms

The algorithm part is the core content of the social influence diffusion part. We take the origin seeding disperse algorithm based on IM model as the baseline algorithm and design a new time algorithm and unequal algorithm to compare with the baseline algorithm. The following sections will cover each of these methods in detail:

Table 4.2: Notation of algorithm

Notation	Definition
G(V,E)	social network graph
g	temporal sub social network graph
V	node of graph
E	edge of graph
S	Seed set
G_t	temporal graph of graph G in time slot t
t	diffusion release time
T	diffusion expire time
S_i	seed set of graph g
$\sigma(S)$	Number of nodes influenced by S with time limit T

4.4.1. Origin seeding disperse algorithm (OSD)

In order to solve time sensitive influence maximization problem and fairness disparity seeding problem, we need to consider the users who are active in time slot t when considering their active patterns. Let $G_t(E_t,V_t)$ be the graph of a social network that only comprises nodes (users) who are active during slot t. $V_t \subseteq V$. After slot t, identify the set of nodes that have previously forwarded the message denoted by C_t. Our algorithm runs through all time slots from t_0 to t_0 + T. For an $edge(u,v) \in E_t$, we allowed node v to pass the message from node u on to node v. Along the time sequence, the nodes that received and forwarded the message will have an impact on their online neighbors. If there are no nodes that are not impacted, or if time T is reached, the procedure finishes. We use simulation to determine if a node is influenced in the process of information dissemination, as shown in Algorithm 1. To see if a node is influenced during the information dissemination process, we employ simulation. In the process of information dissemination, $G_t(E_t,V_t)$ is formed. In time slot t, $G_t(E_t,V_t)$ is produced using nodes and edges. The first graph is a social networking site. A user is represented by a node in the graph. The forwarding times for v to forward u's message are represented by the number on the edge between u and v.

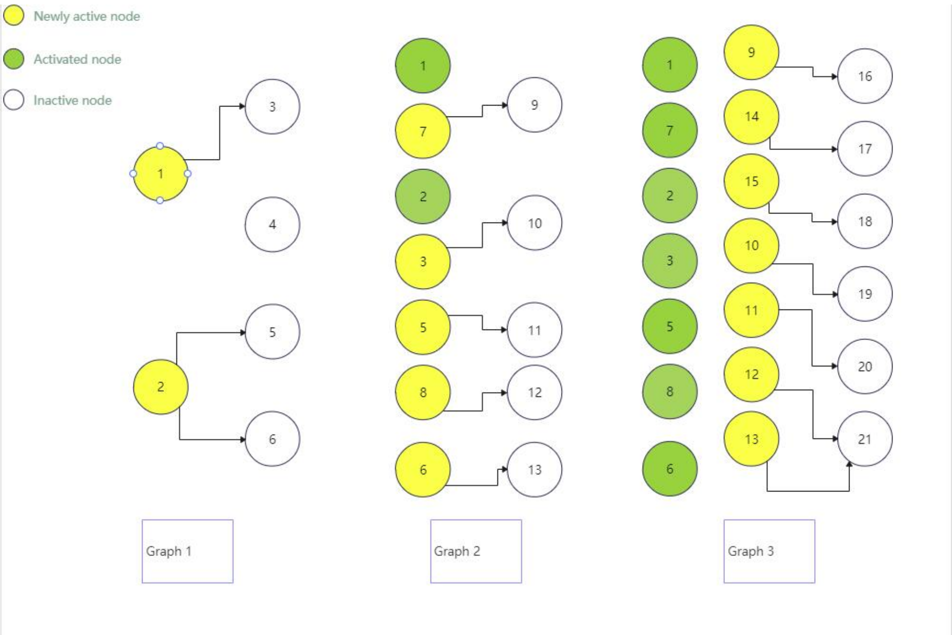


Figure 4.2: The time-awareness diffusion process.

For instance, the number on the edge between V_1 and V_3 is 3, it means that V_3 has interact 3 times to forward V_1's message. The numbers in the brackets near the node is the activities pattern of the user. For example, the active pattern of the user V1 is [0,1,0], it means that the user is only active in the second time slot. The graph of the first time slot shows that only V_1 and V_2 are activities and the edges contain inactivates nodes are pruned. The graph of other time slot is generated in the same way. From fig 4.1, we can see V_1 releases a message (Suppose that if the user is inactivates in the present time slot, it can be forced to become active user if it receive message from V_1) and only V_3, V_4 and V_5 are activity. From Fig 4.1, we can see V_1 releases a message(Suppose that if the user is inactive in the present time slot,it can be forced to go active if it needs to be a source of information) and only V_1 and V_2 are active. At graph graph 1, active user 1 and user 2 send message to user 3 , user 5 and user 6. At the graph 2, It is possible for V_7, V_3, V_5, V_8, V_9 to forward to V_9, V_{10}, V_{11}, V_{12}, V_{13}, the process of information diffusion terminates in the second time slot. And at that time slot, V_1 and V_2 already become activated node and won't being activated again. Then in the second time slot, the new edges in the graph are those between the active users and the inactive users who just receive from the newly actively users. Edges are created between users who forwarded messages in the first time slot and those who are online in the second time slot. The edges that forwarded messages successfully in the first time slot are deleted. If there are no nodes that are not impacted or time T is reached, the procedure ends. The number of influenced nodes is returned when the process completes.

Origin seeding disperse algorithm is an improvement of the original algorithm. Origin seeding disperse algorithm adds the concept of time and improves based on the original algorithm's graph structure. For the weakness of the origin seeding disperse algorithm, due to the previous analysis [52] based on a static graph. However when we consider the temporal effect, the original graph is divided by time, which will generate some separate graph for each time slot. Each graph is different in size and the activeness. But when we distribute seeds for the origin seeding disperse algorithm, we will distribute all seeds to the first graph. The problem is that the selected seed may not appear in the first graph, resulting in inefficient seed propagation. Therefore, according to this situation, we are looking for a better solution that can solve the seeding allocation issues.

Algorithm 1 Origin seeding disperse algorithm

```
$Input$ : G,$t_0$,T,M,P
$Output$ : $\sigma(S)$
for $i = 0 \rightarrow sampling$ do
    for $ratio = 0 \rightarrow ratio[size]$ do
        for $seed = [0] \rightarrow seed[size]$ do
            while $t < t_0 + T$ do
                Generate $G_t(E_t,V_t)$;
                //The edges between the users that have forwarded the message and those active are added
                function ORIGINAL_ALGORITHM($sampling, g, gen, dis, rho$)
                end function
            end while
        end for
    end for
end for
function ORIGINAL_ALGORITHM($sampling, g, gen, dis, rho$)
    for $i = 0 \rightarrow sampling$ do
        for $user \leftarrow seeds.begin()$; $user \rightarrow seeds.end()$ do
            while $diffusion$ do
            end while
        end for
    end for
    $spread_m / = sampling$
    $spread_f / = sampling$
    return $spread_m, spread_f$
end function
```

4.4.2. Equal seeding disperse algorithm (ESD)

We propose Algorithm 2 to simulate the time-awareness influence diffusion along the time sequence and solve the time-awareness influence maximization problem. It is noted that the solution for ESD is differs from the algorithm for the traditional influence maximization problem, because the traditional IM model does not take the time factor into account.

Algorithm 2 (ESD) is based on IM model algorithm. The algorithm diffusion part some of them are likely with the original one, but there is difference. First, we distributed the seeds equally on each graph, and second, we used the time-awareness disparity seeding framework to ensure that our seeds were screened as dynamically as possible. And in order to reduce the computation error, we perform R (i.e. 1000 times) iterations and use the convergence spread influence result to determine which node should be picked as a seed node. R is a given number of simulated information diffusion times. ESD stops when the total budget is reached. Given social network G, the initial time T0, the total time T, P is the number of periods. Based on the diffusion time limit, the number of influencers is generated in social network G through the following steps.

1. First, select the specific ratio value (for example 0.5) from the ratio set R, and select the appropriate seed s from the seed set S which $s_i \in S$, which S is the optimal seed user set of period i.
2. After that, the equal seeding disperse algorithm is called iteratively, and the calculation is carried out following different segments of time.
3. This algorithm is mainly used to calculate how many users are affected by the seed users who are selected by our framework in each period.
4. In the new period (the corresponding graph is generated for each period), we first check the status of activated users in the previous stage as the new part of seed users. Add all last stage activated users in this period as seed users to the seed users set for this current graph, and start marking inactive users in the previous period as active users.
5. The next stage of this algorithm is that we filter users from the seed file based on the target gender ratio and add them to the seed users set.
6. We will do the diffusion by the seed users set in the graph. This step is used to test how many users are activated in this period. After that, the current activated users in the current graph and the previous activated users set would be matched, then select the currently active users who are the inactive users in the previous period to the prepare seed collection, as the seeds users for the next period.

Compared with the baseline algorithm, we can find that the equal seeding diverse algorithm employs an average seed allocation strategy and optimizes the seed allocation for equally seeding diverse algorithms in the early stage but not in the later period. The selection algorithm can improve some shortcomings of the baseline algorithm.

Algorithm 2 Equal seeding disperse algorithm

```
$Input$ : G,$t_0$,T,M,P,seed_limit
$Output : \sigma(S)$
for $i = 0 \rightarrow sampling$ do
    for $ratio = 0 \rightarrow ratio[size]$ do
        for $seed = [0] \rightarrow seed[size]$ do
            $seed\_graph = seed\_limit/seed\_graph$
            for $seed = [0] \rightarrow seed[size]$ do
                // it_seed is the seed that was affected in the previous graph and activated for the first time
                // and seed limit is the total number of seeds, the seed graphs is the number of seeds per graph,for the equal seeding disperse algorithm, the seed graph is same for each graph
                function EQUAL SEEDING _ALGORITHM($sampling, g, gen, dis, rho, it\_seed, seed\_graph$)
                end function
            end for
        end for
    end for
end for
function EQUAL SEEDING _ALGORITHM($sampling, g, gen, dis, rho, it\_seed, seed\_graph$)
    for $i = 0 \rightarrow sampling$ do
        while the number of it_seed is not reached do
            // check it_seed nodes to make sure no active nodes being active by twice time,
            // and filter all the nodes which does not meet the standard.
        end while
        while the number of seed is not reached do
            if user is male and current male number < male limit then
                $s \leftarrow S - 1$
            else
                // user is female and current female number < female limit
                $s \leftarrow S - 1$
            end if
        end while
        for $user \leftarrow seeds.begin(); user \rightarrow seeds.end()$ do
            while $diffusion$ do
            end while
        end for
    end for
    $spread_m/ = sampling$
    $spread_f/ = sampling$
    return $spread_m, spread_f$
end function
```

4.4.3. Unequal seeding disperse algorithm (USD)

Algorithm 1 is a simulation-based influence computation under the IM model without reality temporal consideration. We then propose Algorithm 2, which is a equal seeding disperse algorithm to solve the time-awareness influence maximization problem. It doesn't have a performance guarantee for each time slot with different graph. We further design an improved algorithm in order to maximize social influence spread-ability.

Algorithm 3 (unequal seeding disperse algorithm) uses graph size factor and clustering coefficient to calculate the weight for each graph's seed size. The main idea of this method is to allocate different numbers of seeds in each different period. When the graph is significant, the number of seeds allocated is relatively large when the number of people interacting is high. The main problem solved by this algorithm is that algorithm 2 cannot divide a reasonable number of seeds according to the situation of the graph. In this way, the seed number of the graph with small and low user interaction intensity is the same as that of the graph with significant and high user interaction intensity. Ultimately, it may have negative affects the influence result of diffusion. $\sigma(S)$ is the expected number of influenced nodes by seed node set S. S_i is the subset of the set S, S_i is determined by the weight of combined clustering coefficient and graph size and times the number of total seed set.

1. First, the seed set S has been created according to the upper limit of the number of seeds. The specific ratio value (for example 0.5) has been selected from the ratio set R. And then, according to our seed user allocation formula, we take the target ratio r and the clustering coefficient factor as our input, we form a set of seed allocation coefficients and then decide the number of seeds per period based on the seed allocation coefficients. and select the appropriate seed s from the seed set S which $s_i \in S$, which S is the optimal seed user set of period i.
2. After that, the unequal seeding diverse algorithm is called iteratively, and the calculation is carried out following different segments of time.
3. This algorithm is mainly used to calculate how many users are affected by the seed users who are selected by our framework in each period.
4. In the new period (the corresponding graph is generated for each period), we first check the status of activated users in the previous stage as the new part of seed users. Add all last stage activated users in this period as seed users to the seed users set for this current graph, and start marking inactive users in the previous period as active users.
5. The next stage of this algorithm is that we filter users from the seed file based on the target gender ratio and add them to the seed users set.
6. We employ a dynamic seeding rebalance strategy for unequal seeding disperse algorithm because unequal seeding allocation is based on the seed allocation formula, which could cause the difference between gender ratios and target ratios. Each period we may have a different seed number. In this situation, seed allocation needs to be adjusted dynamically to ensure that our seeding Ratio is the same as the target ratio.
7. We will do the diffusion by the seed users set in the graph. This step is used to test how many users are activated in this period. After that, the current activated users in the current graph and the previous activated users set would be matched, then select the currently active users who are the inactive users in the previous period to the prepare seed collection, as the seeds users for the next period.

Algorithm 3 Unequal seeding disperse algorithm

```
$Input$ : G,$t_0$,T,M,P,seed_limit_set
$Output$ : $\sigma(S)$
for $i = 0 \rightarrow sampling$ do
    for $ratio = 0 \rightarrow ratio[size]$ do
        for $seed = [0] \rightarrow seed[size]$ do
            $seed\_graph = seed\_limit\_set[time\_period]$
            for $seed = [0] \rightarrow seed[size]$ do
                // it_seed is the seed that was affected in the previous graph and activated for the first time
                // and seed limit is the total number of seeds, the seed graph is the number of seeds per graph, in this case the seed graph of per graph is different
                function UNEQUAL SEEDING _ALGORITHM($sampling, g, gen, dis, rho, it\_seed, seed\_graph$)
                end function
            end for
        end for
    end for
end for
function UNEQUAL SEEDING _ALGORITHM($sampling, g, gen, dis, rho, it\_seed, seed\_graph$)
    for $i = 0 \rightarrow sampling$ do
        while the number of it_seed is not reached do
            // check it_seed nodes to make sure no active nodes being active by twice time,
            // and filter all the nodes which does not meet the standard.
        end while
        function GENDER RATIO REBALANCE ALGORITHM($it\_seed, seed\_graph$)
        end function
        //calculating the new graph seeding ratio based on the currently activated users gender ratio, and generate the new male and female limit
        while the number of seed is not reached do
            if user is male and current male number < male limit then
                $s \leftarrow S - 1$
            else
                // user is female and current female number < female limit
                $s \leftarrow S - 1$
            end if
        end while
        for $user \leftarrow seeds.begin()$; $user \rightarrow seeds.end()$ do
            while $diffusion$ do
            end while
        end for
    end for
    $spread_m / = sampling$
    $spread_f / = sampling$
    return $spread_m, spread_f$
end function
```

4.4.4. Comparison of different algorithms

Table 4.3: Comparison of algorithm

Statement and trade-offs		
unequal seeding disperse algorithm	equal seeding disperse algorithm	origin seeding disperse algorithm
+ Can base on graph size and the intensity of graph to allocate seeding adoptively	+ Considering the time factor for seed distribution	- Hard to consider the time factor
+ Can dynamically re-balance the seed ratios	+ seeding ratio for each graph is equal	+ seeding ratio is as the same as gender ratio
+ Consider the re-selecting seeds problem, seed users would not be select twice	- can not dynamically allocate seed based on the characteristics of graph	- No seeding strategy and can not suit for temporal social network

Based on the described in the above algorithm table, the proposed algorithm is being analysed. We design a relatively simple algorithm (basedline algorithm), and it allows us to use the IM model for experimental testing while considering the time factor, but the influence performance is not good compare with the other algorithms. Note that the all seeds allocate at the beginning is not an ideal allocation method in this case, because it has the seed users may not exert their best influence due to not being in the peak influence period, and some selected seed users are not active in the current period problem.

Furthermore, the equal seeding disperse algorithm to be designed to fix the seed allocation problem. This method equally allocate seeds in each time slot, the baseline method has been optimized for seed allocation. Still, this method also has the problem that seeds cannot be allocated according to the size and activeness of the current graph. In this case, it is our intention to design a time-awareness disparity seeding algorithm to overcome the seed allocation problems.

Therefore, in the unequal seeding disperse algorithm designing phase, we need to consider each period might have a different size graph and different activeness, so the seed allocation formula is to be developed. By combining the clustering coefficient and size of graph together, we can allocate the seeds based on the characteristics of the period. And unequal distribution of seeds will cause the gender ratio of seed users to deviate from the target ratio in each period, and we will use seed rebalance to balance the gender ratio dynamically.

5

Implementation and Data Analysis

This chapter describes the experimental setup to answer the research questions by using approaches described in Chapter 4. Some preliminary data analysis is also given, which is not directly relevant to addressing the study questions but provides empirical proof of why the newly designed algorithm is better suited for time awareness than the old method.

5.1. Dataset

Some properties of the dataset are described in this section, together with necessary statistical information, to describe some intuitions about the data distributions and their relevance to our research concerns.

5.1.1. Descriptions

The dataset consists of anonymized user actions on the Facebook and Instagram platform within several years. Each entries contains information on : edgeType, actorId, actionTime, postOwnerId, OwnerFbid, nodeId and female. Explanations on each columns are described in Table 5.1 and 5.2. The dataset is large enough and deemed suitable for use in this thesis. And more statistical details about the dataset have already been described in chapter 3.

In this study, we discuss the Instagram dataset that we used. [50] gathered data on users' reactions to each other's posts through likes and comments, in four years. The dataset was obtained via the Instagram API by first gathering a collection of users, starting with Instagram's founder, and then retrieving the list of followers recursively. Following that, data on the perspectives of the posters, as well as how such posts earn likes and comments from other users, was gathered for several years. The user ID, gender (derived from user names), response type, and other datatypes are recorded for each interaction. A maximum of 5 interactions per post were sampled due to space and processing restrictions. As a result, the dataset's interaction intensity is reduced. The observation period we sample in this paper covers continues 12 months, starting precisely when the number of active Instagram users surged. Before filtering, the dataset contains roughly 8 million distinct individuals with around 38 million interactions between them, based on both likes and comments. We look at the directed network generated by links indicating users liking or commenting on another user's post for both forms of interactions.[52].

We also using facebook dataset to analysis our research question. We gathered data from users at 25 university departments using the Facebook API. The users are 1870 senior students from the aforementioned departments who volunteered to participate, and all conversations between them are recovered. The user ID, gender (as determined by the questionnaire), interaction type, and timestamps are all recorded for each interaction. We also used surveys to obtain user profiles, such as academic standing and hometown. Iterations took place from March 2008 to May 2016, with 97.26 percent of interactions occurring after August 2012. [52].

Table 5.1: Edge description

Variable label	Note	datatype
edge_type	like & comment	string
actor_id	Person ID	string
action_time	Because of data restriction, action_time not always equal to created time.	POSIXct
post_owner_id	The person who published this post.	string
OwnerFbid	The person whom this data was scripted from. Most of time, post_owner_id==OwnerFbid.	string

Table 5.2: Node description

Variable	Node	datatype
node_id	The unique post_id or actor_id, must not be duplicated in node data	string
female	(only available if node_id is person data) 1=female,0=male	integer

5.1.2. Preprocessing

The users in the dataset with non values are filtered so that thoee who interact less than 1 time are removed. Such data filtering strategy is very normal in the influence maximization issues for inactive users. We transfer the raw data to the designed format, as the table 5.1 showing. For the Facebook dataset, We gathered data from users at 25 university departments using the Facebook API. For the INS dataset, we are collecting the collection by using Instagram API, the observation period is from 2015 to 2018.. The collection method and data characteristics is as same as the chapter 3, Facebook and Instagram collection method. In general, in terms of the size of data, the dataset is large and representative fair enough to conduct our research study.

5.2. Experiment

This section explains the detail setup for each task carried out to answer the study questions. Based on the choices of algorithms that are described in Chapter 3, one types of baseline algorithms are implemented, namely original algorithms. And two novel developed algorithm equally seeding and unequally seeding algorithm to be implemented as matched algorithm. Due to this thesis is to undertake with the last project[52], there are seveal programming language to perform all necessary steps, from data analysis to the implementation of the algorithms, is needed. The MYSQL to be used to create the raw data file for the input file of the preprocess stage, MySQL is offered under a variety of proprietary licenses and free and open-source software under the rules of the GNU General Public License. MySQL was developed and supported by MySQL AB, a Swedish business that was later purchased by Sun Microsystems (now Oracle Corporation). And MYSQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database. [44]. It is decided that Python is used because it is currently the most popular language for data science. Python to be used to preprocess the data in this thesis, because NumPy, SciPy, and Matplotlib are examples of Python libraries that can be used in scientific computing, Its library includes algebra, combinatorics, numerical mathematics, number theory, and calculus, among other topics. [43]. C++ also to be used to implement the diffusion algorithm because based on [51] C++ was created with a focus on system programming, embedded, resource-constrained software, and large systems in mind, with performance, efficiency, and flexibility as design goals. C++ provide us a faster running speed to run the diffusion algorithm code with lower compile time, It also offers convenient methods to perform standard evaluation strategy such as metrics calculation. This library is mainly used to perform experiments on the data formatting and scientific computation task.

The data preprocessing and evaluation part is implemented by Python. A standalone Python script is utilized to do the evaluations. However, to preserve strong repeatability of the experimental setup, it was chosen to implement most of the data pipelines in Mysql.

5.2.1. Set up

The demo system of Disparity Seeding is on the github website [1] . We compare four variants of Disparity Seeding (i.e., Target HI-Index, indegree, intensity, pagerank) with the state-of-the-art

[1] https://github.com/123AB/OnlyDiffusion

approaches: time awareness diversity seeding algorithm and IM-balanced-enhanced. Both time awareness diversity seeding algorithm and IM-balanced are not intended to achieve the target gender ratio ζ. The users' interaction is divided by month into 24 intervals over time, each interval is one month in length, and 24 intervals are continuous. We use diversity seeding and IM-balanced to keep track of the target gender ratio for each time interval. Time awareness diversity seeding chooses a seeding ratio between ζ and the gender ratio of the top K users with the highest in-degree, maximizing the influence within each time interval. IM-balanced greedily chooses seeds to increase influence while guaranteeing that female influence is at least ζ equal to the optimal female influence for each time interval. The diffusion is then simulated using the Independent Cascade diffusion model [26], with the likelihood of user v_i influencing user v_j set to the number of likes/comments v_i gives to v_j divided by the total number of likes/comments received by user. Where the simulated diffusion is continuous run per time interval by time interval until all of the time interval being finished.

With varied goal gender ratios, we examine the spreading gender ratio and influence spread. The performance measures are: 1) the absolute error, i.e., |s - k|, where s is the spreading gender ratio, and 2) the spread of influence. On the Facebook, the seed group size K is set from 200 to 50, specificly which are 200, 175, 150, 125, 100 and 50. Respectively, and the error margin is set to 25% on both datasets. The size of seed groups in Phase II of Disparity Seeding has the same setting as the pervious one. On The Facebook, respectively. The results of each simulation are averaged run 1000 times. The diffusion process tests are carried out on an Dell G15 server with an Ryzen 5 5600H 3.30GHz processor and 8GB RAM, the disparity seeding perform the test on 16 Cores, 64GB RAM google cloud standard machine.

On the Instagram, the seed group size K is set from 200 to 50, and as well as 5000. The size of seed groups in Phase II of Disparity Seeding has the same setting as the pervious one. On The Instagram, respectively. The results of each simulation are averaged run 1000 times. The diffusion process tests are carried out on an Dell G15 server with an Ryzen 5 5600H 3.30GHz processor and 8GB RAM, the disparity seeding perform tests on an google cloud platform server with an intel broadwell 16G processor and 64GB RAM google cloud standard machine.

5.2.2. Implementations of the seeding algorithms

Before running the diffusion, we need to create seeding files for each set of different indexes. We filtered missed and non-standard data and generated corresponding seeding files according to different time intervals and ratios. In this experiment, the ratio set is from 0.0 to 1.0 and seed set is include 50,75,100,125,150,175,200. Complete permutations are formed to produce different seed file combinations. And then, we rank the users based on each difference index for each time slot. After the user ranking stage, the seeding files would be generated based on user ranking file of each time slot. It is essential to emphasize that for the unequal seeding disperse algorithm. We also need to calculate the seeding factor based on the seeding allocation formula that we developed in chapter 4 and use this factor times the total number of seeds. The result is the input parameter of the unequal seeding disperse algorithm for each period.

5.2.3. Simulation result on Facebook

In this section, the total outreach analysis to analyze which algorithm can maximize social influence considering the time factor. The absolute error analysis to analyze which algorithm has the minimum error. the Feasibility analysis to analyze Which algorithm has the better applicability, which means the winner algorithm performance best in different set seeds. And then, the seeding and influence analysis to analyze the relationship between the number of seeds allocated and the number of influenced users helps us better understand the characteristics of user activity patterns when considering the time factor. Lastly, the sensitivity analysis to analyze The impact of different time periods on the number of influenced users.

Total outreach analysis

Table 5.3: Facebook dataset with Ratio 0.5 Seed 100 during 24 months.

algorithms	Index (like)			
	Target HI-index	Indegree	Intensity	Pagerank
origin seeding disperse algorithm	466.60	265.41	407.33	438.74
equal seeding disperse algorithm	667.34	670.03	631.15	602.21
unequal seeding disperse algorithm	774.22	816.15	811.23	796.21

Table 5.4: Facebook dataset with Ratio 0.5 Seed 100 during 24 months.

algorithms	Index (comment)			
	Target HI-index	Indegree	Intensity	Pagerank
origin seeding disperse algorithm	425.46	304.41	437.24	415.49
equal seeding disperse algorithm	521.60	602.87	510.83	507.94
unequal seeding disperse algorithm	647.55	712.81	652.23	706.21

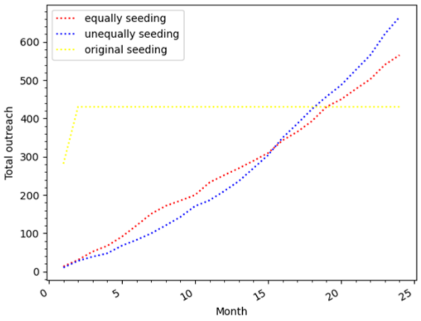


Figure 5.1: The total outreach of the comment using target HI-index.

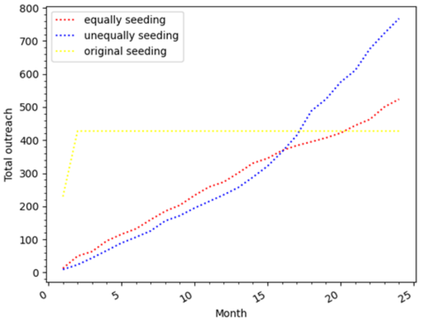


Figure 5.2: The total outreach of the like using target HI-index.

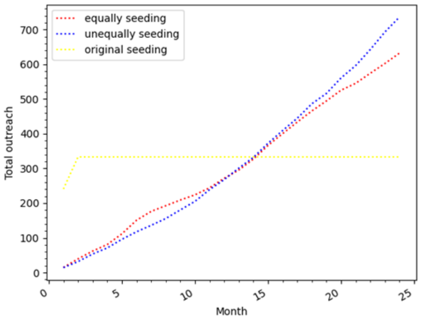


Figure 5.3: The total outreach of the comment using indegree.

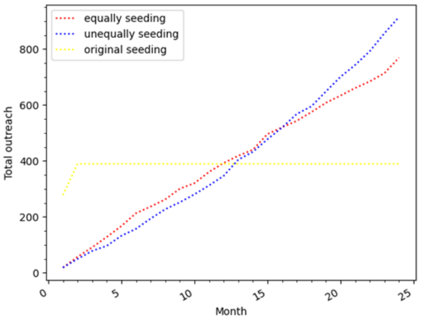


Figure 5.4: The total outreach of the like using indegree.

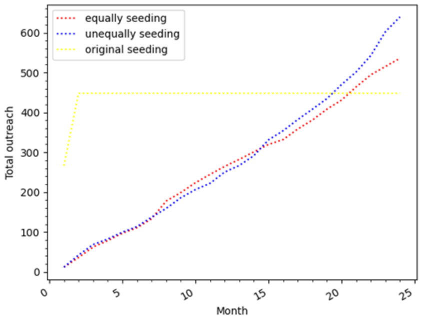


Figure 5.5: The total outreach of the comment using intensity.

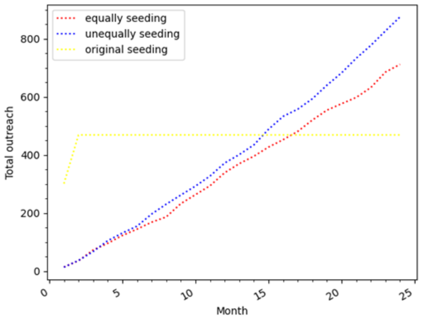


Figure 5.6: The total outreach of the like using intensity.

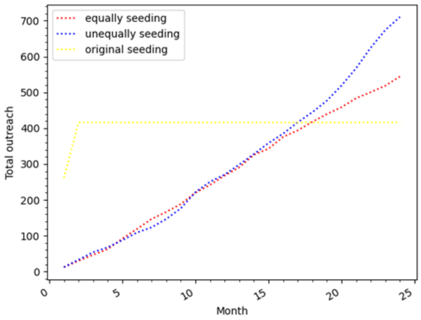


Figure 5.7: The total outreach of the comment using pagerank.

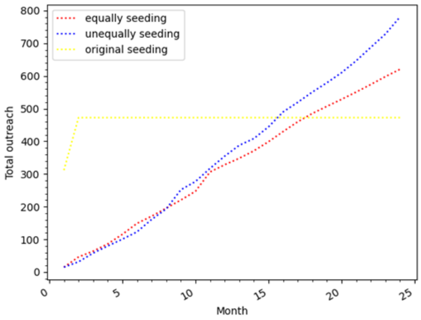


Figure 5.8: The total outreach of the like using pagerank.

Our first set of experiments is conducted over all sessions in the test set and related to answer **research question 2**: *Can we improve the disparity seeding algorithm by accommodating temporal effect in social networks to reach fairness of gender ratio for seeds selection?* The result of the Facebook dataset is represent in tablex.

For the total outreach analysis, we are use the seed 100, 24 months period with target ratio 0.5 as our sample data. table 5.3 show the total outreach result for each index, target hi index, indegree, intensity and pagerank. The target hi index in terms of like tag for unequally seeding diverse performance 16% higher than the equally seeding diverse, and 66% higher than the basedline seeding diverse algorithm. Similarly, the target hi index with comment tag performance performance 25% higher than the equally seeding diverse, and 55.8% higher than the baseline seeding diverse algorithm. This seeding method presents good advantages compared with the equally seeding diverse algorithm and the baseline algorithm.

From the figure 5.1 to figure 5.8, we can see that the x-axis represents the period, the y-axis represents the number of influenced people, and the blue line, red line, yellow line represent unequal seeding disperse algorithm, equal seeding disperse algorithm and origin seeding disperse algorithm. These figures show that the baseline algorithm showed an early advantage because our baseline algorithm seeding strategy focused on whole seeds allocation in the first month of the 24 months experimental period.

However, the baseline seeding strategy did not perform very well in the later period. The number of affected users gradually decreased due to the lack of subsequent seed users to join in. The equally seeding diverse algorithm It has an advantage for unequally from early to middle experimental period, because our equally seeding strategy allocates seeds equally, our social network graph has a relatively small graph size and lower user interaction activities. Therefore, the number of seeds allocated to unequally seeding diverse algorithm is less than equally seeding diverse algorithm. So there is an early lead for equally seeding diverse algorithm to unequally seeding diverse algorithm. However, because the graphs gets bigger on later period and users are more active which means the graphs have higher clustering coefficient compare with the earlier graph. Then we can see the unequally seeding diverse algorithm have better performance compare with the baseline algorithm and equally seeding algorithm at the later half experimental period.

Based on our observation and analysis of the data, we find that the number of influenced persons remains low at the beginning and has a higher number after the mid-period. The algorithm performs better than the other two algorithms regarding the number of affected persons.

Absolute error analysis

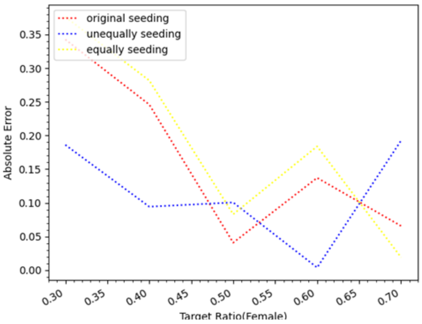


Figure 5.9: The absolute error of comment using the target hi index.

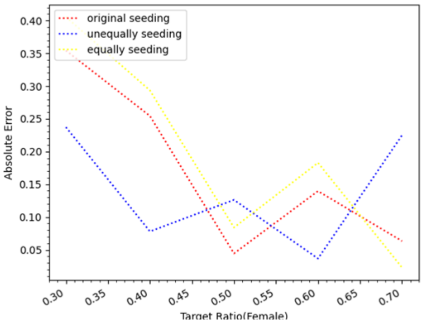


Figure 5.10: The absolute error of like using the target hi index.

Our second set of experiments is conducted over all sessions in the test set and related to answer **research question 3**: *Can we improve the disparity seeding algorithm by Utilizing temporal effect in social networks to achieve the goal of influence maximization?* The result of the Facebook dataset is represent in figure 5.9 and 5.10.

To analyze which algorithm could provide less absolute error, we will conduct the absolute error analysis. The definition of absolute error means The degree to which influenced ratio deviates from target ratio. Therefore the absolute error = influenced ratio - target ratio. We are using the target ratio as our first input ratio. We will get the seeding ratio after the diffusion process and scaling function process. Then we will take the seeding ratio as our input ratio to generate influenced ratio. The absolute error would be the difference between influenced ratio and target ratio.

From figures 5.9 and 5.10 we can see that manifest the absolute errors (y-axis) under Time awareness Disparity Seeding with Target HI-index and diversity seeding with a varying target gender ratio (x-axis). It's worth noting that a strategy with a lower absolute error is more likely to meet the desired target gender ratio requirement. We have two points to concluded. Disparity Seeding has the least absolute errors for most of the target gender ratios. Disparity Seeding carefully selects the seeding ratio by capturing the relationship between the target and seeding ratios, but there is no clearly relationship between seeding ratios and our influential ratios. Target HI-index considers the target ratio in ranking and penalizes users who do not meet the requirement, allowing for greater flexibility in accommodating extreme target ratios. Based on our test result, the target HI-index perform best among the other indexes to reduce the absolute error. On the other hand, Diverse seeding uses the target ratio as an input. Still, it only seeks a ratio that maximizes the influence spread rather than minimizes the discrepancy between the spreading and target ratios. Second, when the target ratio is between 0.5 and 0.65, the absolute errors are the smallest for both approaches. Because the female ratio on Facebook is 58.97 percent, both ways are more likely to meet the target gender ratio of roughly 50 to 60 percent.

The unequal seeding disperses algorithm has fewer absolute errors than the other two algorithms, whether in the gender ratio close to the ratio of original distribution or deviating from the initial distribution.

Feasibility analysis

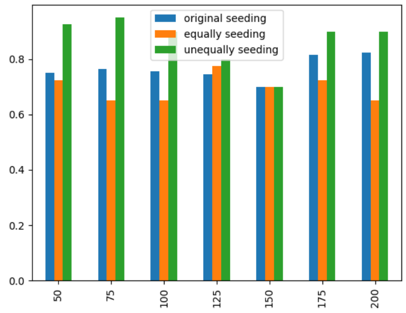


Figure 5.11: The feasibility ratio of the comment using target HI-index.

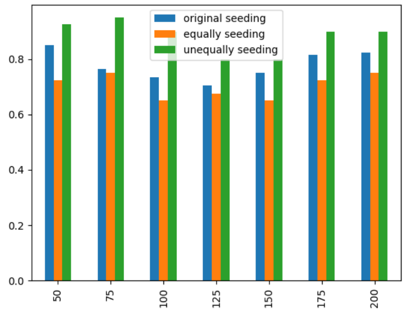


Figure 5.12: The feasibility ratio of the like using target HI-index.

In order to compare the absolute error of different algorithms, the feasibility analysis was adopted. Based on [23] a feasibility study determines if a project or system is feasible. A feasibility study tries to objectively and logically identify the strengths and weaknesses of a current model, the possibilities and risks that exist in the natural environment, the resources needed to carry out the project, and the likelihood of success. In our project, the feasibility means for different ratios of the number of seeds in the same set, the number of the group passed the margin errors divided by the total number of experiments. The experiment set the margin error to be 0.20. We selected 11 random target ratios as the input then used the influenced ratios of the 11 input ratios as the output. Then, the absolute error was obtained by manipulating influenced ratio - target ratio. Then we compared absolute error with margin error. We accept it if it is smaller than margin error; if it is more significant than margin error, we reject it. After that, we divide the number of all accepted ratio experimental groups by the total number of experimental groups.

From the figure 5.11 to figure 5.12, we can see that the x-axis represents the difference number of seed users, the y-axis represents feasibility ratio, and the green bar, red orange bar, blue bar represent unequal seeding disperse algorithm, equal seeding disperse algorithm and origin seeding disperse algorithm.

According to the feasibility, unequal seeding disperse had the best performance among the three seeding algorithms. This means that our algorithm has better stability and applicability. Because our algorithm adopts dynamic seed re balance, it ensures the balance of seed ratio at any time. It ensures that the seed ratio will not deviate from the target ratio as the number of seeds increases.

The unequal seeding disperses algorithm has better feasibility ratio than the other two algorithms over all of the seed sets.

Seeding&influence analysis

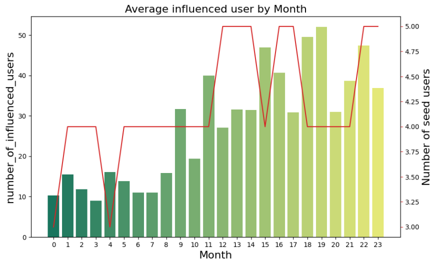


Figure 5.13: The average influenced users by month of the comment using target HI-index.

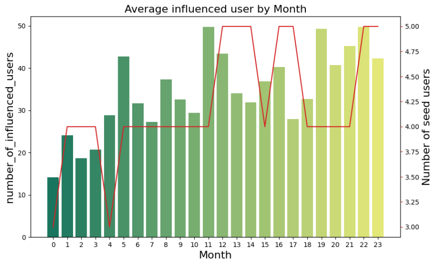


Figure 5.14: The average influenced users by month of the like using target HI-index.

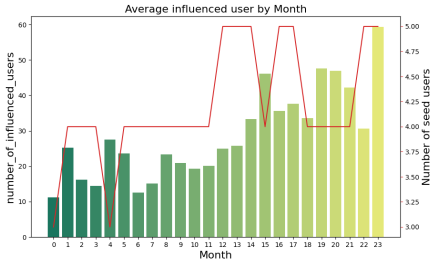


Figure 5.15: The average influenced users by month of the comment using indegree.

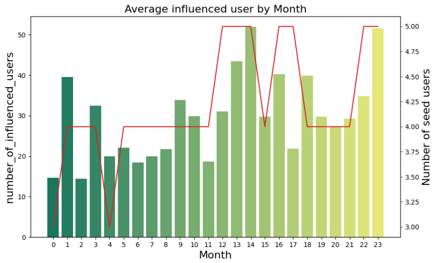


Figure 5.16: The average influenced users by month of the like using indegree.

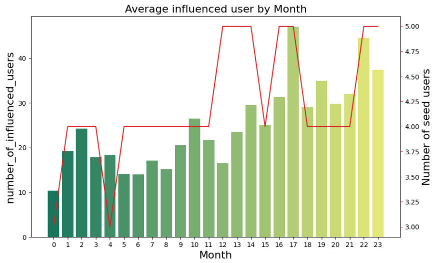


Figure 5.17: The average influenced users by month of the comment using intensity.

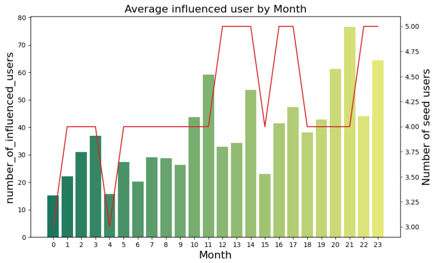


Figure 5.18: The average influenced users by month of the like using intensity.

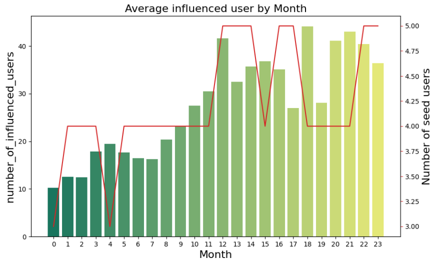


Figure 5.19: The average influenced users by month of the comment using pagerank.

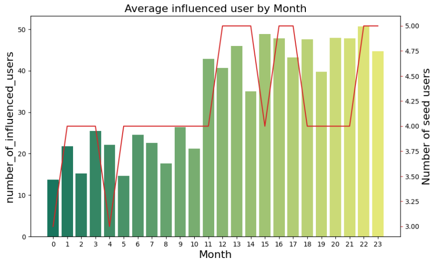


Figure 5.20: The average influenced users by month of the like using pagerank.

The seed influence relationship analysis adopts to explore the relationship between the number of seed users and the number of affected people. We are using the seed 100, 24 months period and target ratio 0.5 as our input factor. From the figure 5.13 to figure 5.20, we can see that the x-axis represents the period, the left y-axis represents the number of influenced people, and the right y-axis represents the number of seed users in each time slot. By observes the number of seed distribution and the number of influenced users have a positive correlation because we can see from the data that when the number of influenced users increases, the number of seeds is higher than the average seed numbers. When the number of influenced users is lower than average, the seed numbers are also lower than the average. From the analysis in the figure, we can also see that different indexes impact the final number of affected people. Still, there is not much difference in the distribution pattern of seed users and the growth pattern of influenced users.

> The number of seed users allocated in each time slot and the number of influenced users has been shown a positive correlation.

Sensitivity analysis

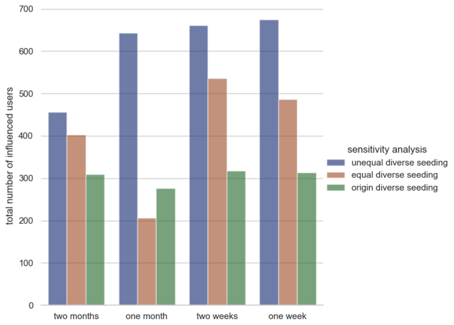


Figure 5.21: Sensitivity analysis of comment using indegree.

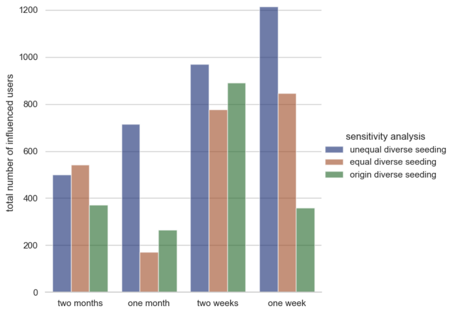


Figure 5.22: Sensitivity analysis of like using indegree.

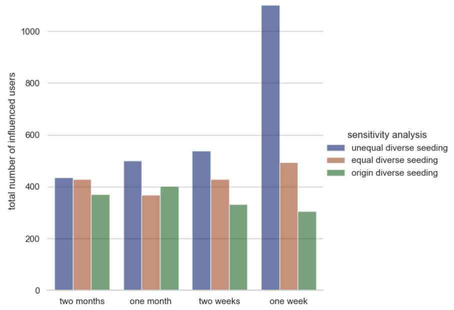


Figure 5.23: Sensitivity analysis of comment using intensity.

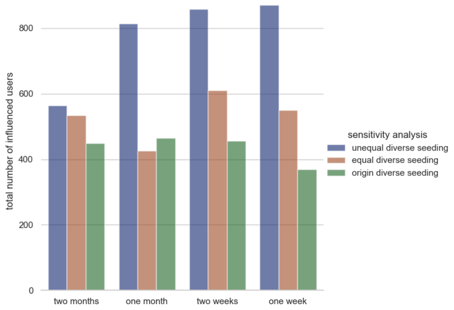


Figure 5.24: Sensitivity analysis of like using intensity.

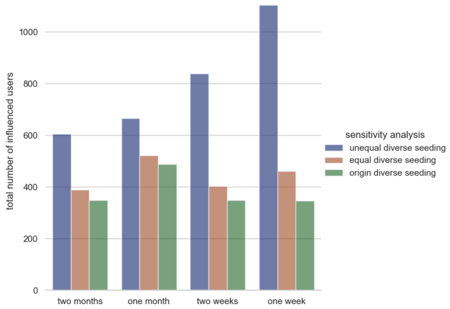


Figure 5.25: Sensitivity analysis of comment using pagerank.

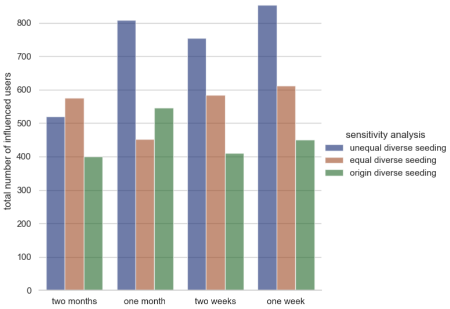


Figure 5.26: Sensitivity analysis of like using pagerank.

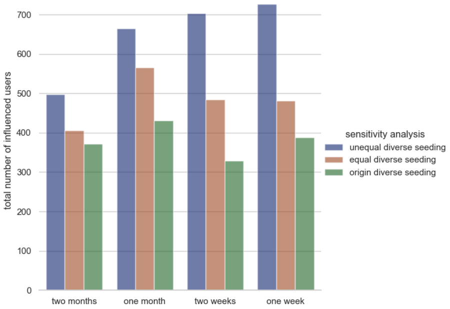


Figure 5.27: Sensitivity analysis of comment using target HI-index.

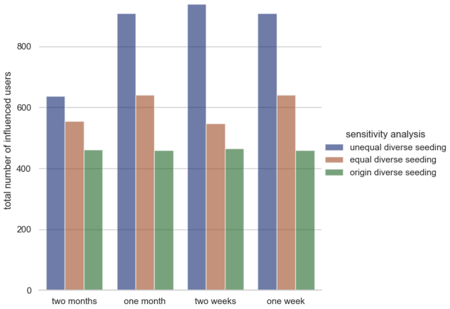


Figure 5.28: Sensitivity analysis of like using target HI-index.

To find out the sensitivity factors that have an essential impact on the influence diffusion, sensitivity analysis has been conduct in this research. Based on [49] mentioned, the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) might be divided and assigned to various sources of uncertainty in its inputs is known as sensitivity analysis.
Therefore, multiple uncertain factors have been considered by analyzing and measuring their impact on the number of influenced users in relation to sensitivity degree, furthermore to prove the credibility of the experimental results.

From the figure 5.21 to figure 5.28, we can see that the x-axis represents the difference period, the y-axis represents the number of influenced users, and the green bar, red orange bar, blue bar represent origin seeding disperse algorithm, equal seeding disperse algorithm and unequal seeding disperse algorithm.

From the data contained in the graph, The number of influenced users by the unequal seeding diverse algorithm is gradually increasing from the period two months to one week. This means that with the smaller interval, the seed we select is the optimal local solution, which can eventually produce a better social influence effect. Because we choose the most influential seed users in a particular period based on the time factor, which means that when spreading, the users we choose are always the most effective.

By comparing different indexes over different time intervals, we find that the number of users influenced by the origin seeding diverse algorithm is not strongly correlated with the interval. However, we can observe that the baseline algorithm performs better when the time interval is longer. This is because we have a larger user interaction graph with a longer time interval, and the baseline algorithm can affect more users at the beginning. The change in the length of the period did not positively affect the origin seeding diverse algorithm.

The target HI-index is our main focus index, and to calculate the sensitivity analysis ratio for this index. We are using the percentage change in total outreach divided by percentage change in the duration of each period. For comment posted by users in terms of unequal seeding disperse algorithm, if we change the time period from one month to two month, the sensitivity ratio is 13.84%, the time period change from one month to two weeks, the sensitivity ratio is 7.7% and the time period change from one month to one week, the sensitivity ratio is 3.08%. That means with the input value duration of time period change, the output value total outreach also changed significantly.For the like posted by users, if we change the time period from one month to two month,the sensitivity ratio is 28.76%, if we change the time period from one month to two weeks, the sensitivity ratio is 34.2%, which increase significantly, and the time period change from one month to one weeks, the sensitivity ratio is 16.10%. we found that the number of influenced users would also change significantly over time by validating the percentages of number of user comments and likes varies with the percentages of period duration change.

From the above figures, we can observe that with the time interval varying, the equal seeding diverse algorithm does not show similar results to those of unequal seeding diverse algorithm. The reason is that as the time interval becomes smaller, more graphs are generated. However, the Equal seeding diverse algorithm does not appropriately adapt to this situation and distributes seeds evenly to different time intervals. Therefore, the number of graphs intensifies the disadvantage of this algorithm compared with the diverse algorithm of unequal diverse seeding algorithm.

With the period change, the unequal seeding disperse algorithm still shows a similar trend among difference periods, reflecting the algorithm's stability and applicability.

5.2.4. Simulation result on Instagram

The experiment of Instagram is mainly used to verify the conclusions drawn from Facebook data and further support our findings on research questions 2,3. Furthermore, Get some insights about the characteristics of user activity patterns in a dynamic social network with big user data.

Total outreach analysis

Table 5.5: The total outreach duration analysis.

algorithms	Index (like)			
	Target HI-index	Indegree	Intensity	Pagerank
origin seeding disperse algorithm	50723.20	51149.3	51523.5	51628.2
equal seeding disperse algorithm	778829.20	751732.40	728596.7	768616.49
unequal seeding disperse algorithm	840044.5	778169.70	783956.10	846729.75

Table 5.6: The total outreach duration analysis.

algorithms	Index (comment)			
	Target HI-index	Indegree	Intensity	Pagerank
origin seeding disperse algorithm	12352.69	12964.56	13027.0	13532.07
equal seeding disperse algorithm	177029.47	175926.72	217226.79	208465.78
unequal seeding disperse algorithm	192161.98	193272.24	225482.8	221344.7

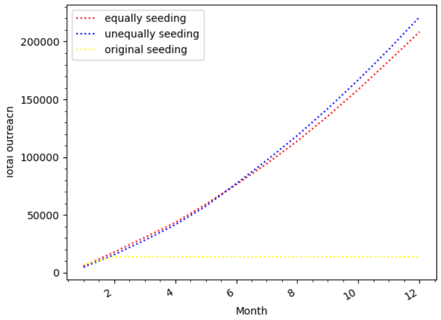


Figure 5.29: The total outreach of the comment using pagerank.

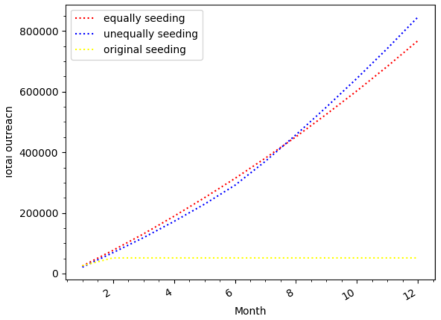


Figure 5.30: The total outreach of the like using pagerank.

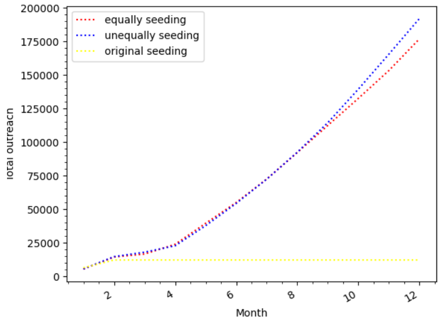


Figure 5.31: The total outreach of the comment using target HI-index.

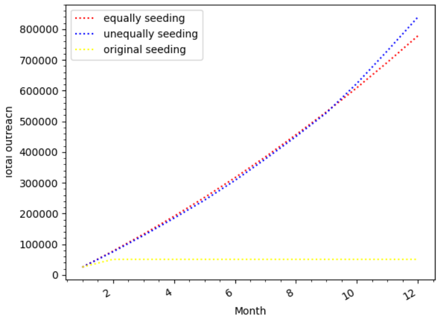


Figure 5.32: The total outreach of the like using target HI-index.

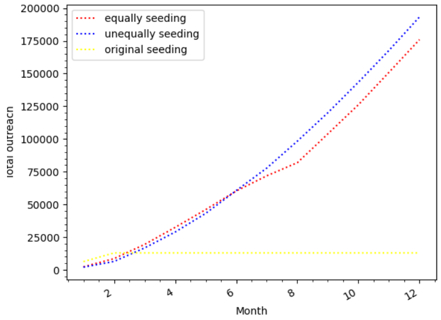


Figure 5.33: The total outreach of the comment using indegree index.

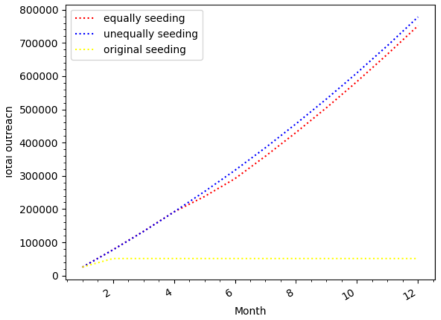


Figure 5.34: The total outreach of the like using indegree index.

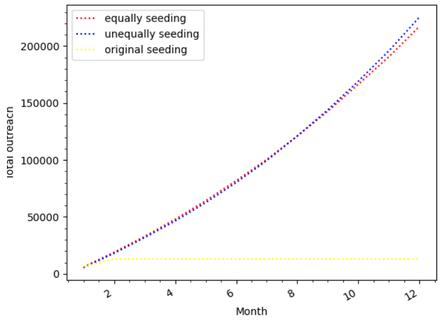


Figure 5.35: The total outreach of the comment using intensity index.

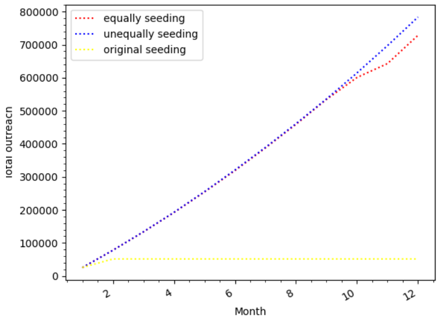


Figure 5.36: The total outreach of the like using intensity index.

For the total outreach analysis of Instagram, we are use the seed 100, 12 months period with target ratio 0.5 as our sample data. From table 5.5 we can see that, unequal seeding diverse algorithm performance 7.86% better than equal seeding, performance 15 times better than the baseline seeding algorithms with the like tag. For the target hi index with comment tag, unequal seeding diverse algorithm performance 8.5% better than equal seeding, performance around 15 times better than the baseline seeding algorithms.

From figures 5.29 to figures 5.36 we can see the total outreach duration analysis of Instagram, the baseline algorithm has the earlier lead at the beginning, but performance poorly compare with the unequal seeding disperse algorithm and equal seeding algorithm. The reason for this is Instagram With a larger database, the disadvantage of the baseline algorithm is that the seed cannot be invested in the subsequent time, resulting in the spread only being carried out in the period when the early stage is relatively small, or even the seed is not active. That will lead to a vast difference diffusion result after several iterations.

The graph size will increase with the time period increase, the original seeding diverse algorithm can not adapt this change very well.

The equally seeding perform a little bit lead at the earlier stage of the diffusion process, however, it is overtaken by the unequally seeding algorithm and even surpassed by the USD algorithm after the mid-stage. Which showing the similar trend with the Facebook dataset. The reason for this is the graph structure of the two datasets has the same changing trend with graph size in terms of temporal effect. Therefore the USD algorithm in the second half of the spread results are better, because more seeds are distributed through unequally seeding diverse algorithm, and more influence is generated. Because the second half of the experimental period has more users than the first half, more users will be affected if more seeds are put into the second half of the experimental period. The performance of unequally does not exceed equally by much due to the high amount of Instagram data. The increase of tens of thousands of users is not obvious in percentage but is significant in absolute number, which is also essential for influence spread.

Absolute error analysis

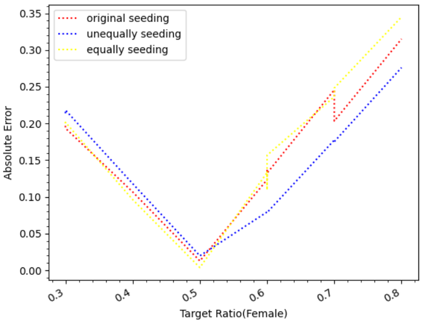


Figure 5.37: The absolute error of comment using target HI-index.

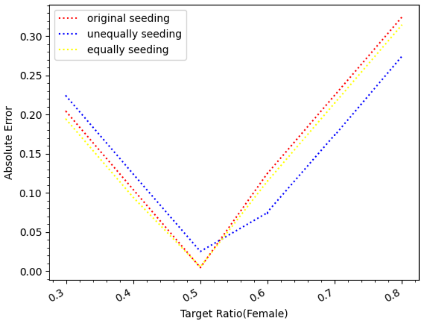


Figure 5.38: The absolute error of like using target HI-index.

The procedure of absolute error analysis is as same as the Facebook test. From figure 5.37 to figure 5.38, we can see that the absolute error at the target ratio of 0.5 is minimal compared with other ratios. The reason is that according to the statistical analysis of Ins dataset in Chapter 3, it is found that the overall influence spread trend is close to this figure because female users account for 58 %. But the 0.1% percent of users women accounted for only 47%. Our research has also found that the influence of the same people tends to have higher interaction among each other. The influence of the higher rank group, the male has a higher proportion, so in the first 100 seed users, the absolute error at target ratio 0.5 is minimum among other ratios is in line with the statistical analysis of the situation. It can be seen from the figure that the absolute error of original seeding diverse algorithm and equally seeding diverse before target ratio at 0.5 is lower than that of unequally seeding diverse algorithms. There is a relatively high number of users using an unequally seeding algorithm, and the performance after 0.6 is better because it is more consistent with the real population distribution. The two algorithms have an advantage in influence diffusion and are easier to control the gender ratio due to the relatively small number of participants.

Feasibility analysis

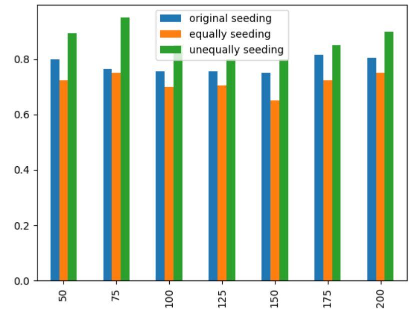


Figure 5.39: The feasibility ratio of comment using the target HI-index.

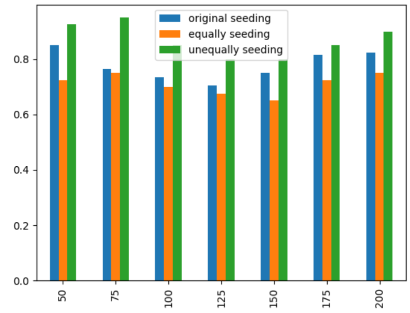


Figure 5.40: The feasibility ratio of like using the target HI-index.

The idea of the feasibility analysis is as same as the Facebook experimental test. From the figure 5.39 and figure 5.40, we can conclude that we have the similar result with the Facebook dataset. The unequal seeding disperse algorithm has the highest feasibility ratio, and origin, equal seeding disperse has the similar feasibility ratio.

Seeding&influence analysis

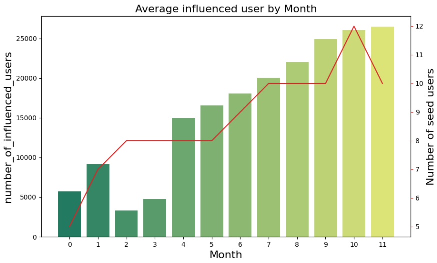


Figure 5.41: The average influenced users by month of the comment using target HI-index.

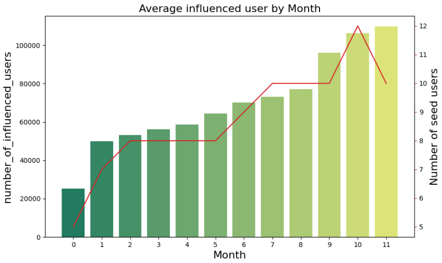


Figure 5.42: The average influenced users by month of the like using target HI-index.

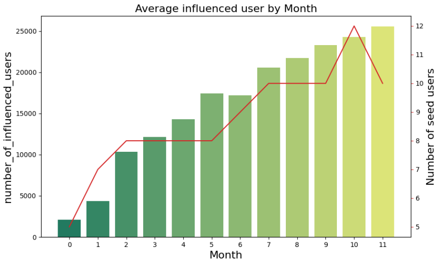


Figure 5.43: The average influenced users by month of the comment using indegree.

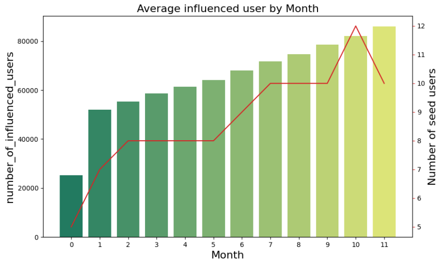


Figure 5.44: The average influenced users by month of the like using indegree.

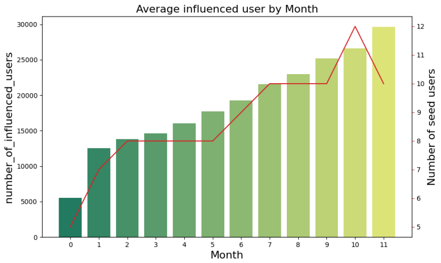


Figure 5.45: The average influenced users by month of the comment using intensity.

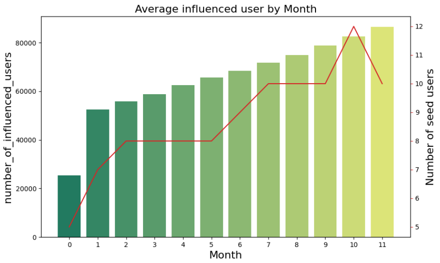


Figure 5.46: The average influenced users by month of the like using intensity.

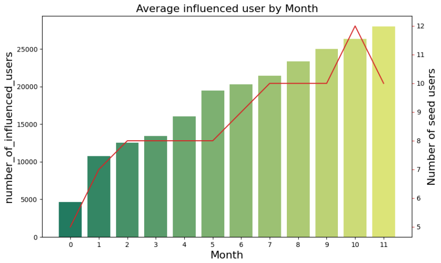


Figure 5.47: The average influenced users by month of the comment using pagerank.

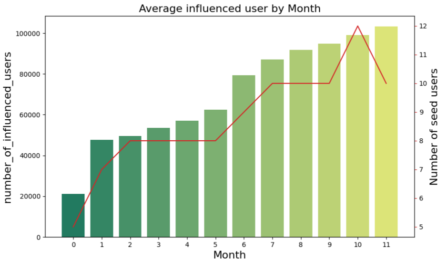


Figure 5.48: The average influenced users by month of the like using pagerank.

From the above figures figure 5.40 to figure 5.48, we can see that the distribution of seed users. The number of seeds is small at the beginning of the diffusion and gradually increases over time. The number of users we influenced also showed a low number of influenced users trend at the beginning of the graph and then a relatively high number of influenced users trend after the mid-period of the graph, basically in line with our seed user trend.

5.2.5. Big seed set analysis

Table 5.7: Instagram total outreach with Seed 5000.

algorithms	Target HI-index	
	like	comment
origin seeding disperse algorithm	52003.10	20759.69
equal seeding disperse algorithm	791602.66	204021.90
unequal seeding disperse algorithm	864158.29	226092.87

Because Instagram has many more users than Facebook, the 50 to 200 seed user sets previously tested were limited in reflecting users' behavior patterns who weren't at the top level of influence. So we decided to use seed user 5000 with target ratio 0.5 as our input parameter. And based on our testing, we can see the result from table 5.7, the total number of influenced people is more than 200, but it doesn't show a percentage increase because our algorithm already covers most of the active users at this time. Then we also calculate the cover rate, We looked at the total number of unique users over 12 months, and it was about 950,000. The total influenced user for seed users 100 is around 840,000, the cover rate is about 88%, the total influenced user for seed users 100 is around 860,000, the cover rate is about 90.5%. Therefore even for big data set, our algorithms also showing a good result.

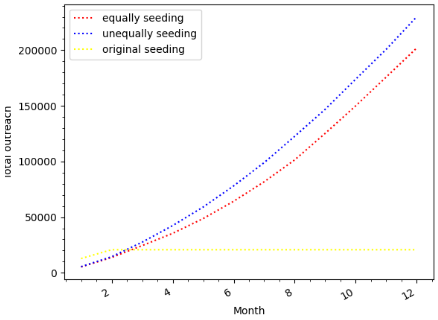


Figure 5.49: The total outreach of the comment using target HI-index.

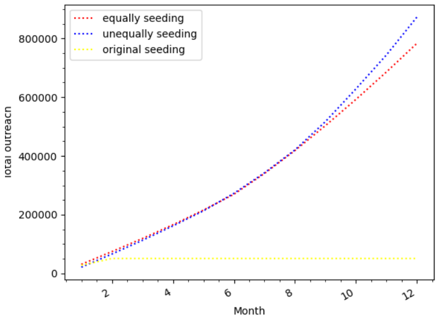


Figure 5.50: The total outreach of the the like using target HI-index.

From the figure 5.49 to figure 5.50, we can see that the setting is as same as the Instagram except the number of seed set, which is 5000. The result also showing the similar trend, this means that data validation through large seed data sets can also prove that our algorithm effectively increases the number of influenced users.

5.2.6. Discussion

The finding presented the potential of using the time-awareness disparity seeding framework to solve the gender equality and influence maximization problem. We have shown that considering the time factor, unequal seeding disperse algorithm is already a good method for seed allocation and influence maximization. By combining clustering coefficient and graph size factors both yields the good seeding allocation model, and we can even improve the performance of influence maximization and reduce the absolute error by using this method.

The empirical result simply illustrates the if we can allocate more seed users to the graphs with more activeness, we can improve our influence distribution efficiency. Among these algorithms that have been tested in this project, unequal seeding disperse algorithm not only solve the problem about influence maximization but also can reduce the absolute error at the same time. Which is also one of the goal of our research. This demonstrates that seed allocation can further reduce absolute error through dynamically seeding by better integrating the Target HI-index method to identify more qualified users after targeted HI-index ranks. Because the effect of target HI-index depends on the gender ratio of the nodes adjacent to the seed nodes. Our dynamic seed allocation allows the seed users to be selected in a suitable time interval.

6

Conclusion and Future Work

The first section of this chapter highlights the conclusions that may be taken from the experimental results presented in this thesis. We also suggest a disparity seeding algorithm that may be useful. Finally, we conclude this thesis by identifying some prospective future research projects based on the findings of this thesis.

6.1. Conclusion

This section reflects on the empirical data that we discovered in our experiments to address the research questions that we identified at the start of this thesis.

> **Research question 1.** *Can we successfully find the behavior patterns and characteristics of users in temporal social networks?*

To answer this question, we had analyze the overall active trend of users and studied the three sub-problems mentioned in chapter 3. We found that the characteristics of active users can be acquired from the temporal social networks.

To identify the basic temporal trend of the active users, we need to analyze the total number of users, the proportion of female and male users, and the overall trend of active users over time. we can initially show the basic phenomenon that the total number of active users is growing rapidly between week 100 up till the week 200 and reach the peak after 200 weeks.

We have also laid out some of the challenges and ideas on how the influence trends change over time for different influence user groups. To explore the homophily effect in the different levels of users, we are researching the social interaction of the homophily groups between males and females. We found that the homophily effect is not strongly apparent between the same gender, however, it is strongly evident between users with similar ranks.

Based on the basic trend of temporal effect, we found that with the increase of users' level of influence, the average interaction also increases. The average indegree of all users is around 7 and 298 for the top 0.1% users. Even though users don't reach the peak of their influence in a short period of time, we need to consider when they do. As a result, we started analyzing the time it takes for top users to reach their most influential week. This helped us find that the more influential the users are, the longer it takes for the them to reach their peak of influence in relation to temporal effect. Based on the rich get richer mechanism, we proposed that disparity of users' activeness and gender may yield different active durations. According to the statistic analysis and after analyzing the question from chapter 3, we conclude that The average intensity of both males and females results are in increased trend. This represents the rich get richer mechanism considering temporal effect.

Overall, Through statistical and quantitative analysis, we summarize the characteristics of users under the temporal factor.

Research question 2. *Can we improve the disparity seeding algorithm by accommodating temporal effect in social networks to reach fairness of gender ratio for seeds selection?*

Based on our experimental findings, by dynamically balancing user gender ratio in the process of seed allocation through unequally seeding diverse algorithms, we found a slight improvement. Nonetheless, the incremental difference is negligible, and more research is needed to confirm the utility of such method.

The gender ratio of seed users may change in the diffusion process, so we dynamically adjust the gender ratio of users to ensure the real-time balance of seed users' target gender ratio. The unequally seeding diverse algorithm helps us build an advantage in the later stages of diffusion because there are usually more active users interaction and more users on the later graph. We can ensure that our overall diffusion is a correction of natural diffusion, thereby reducing the error after target ratio 0.5 and establishing an advantage over the first two algorithms. The ratio of natural diffusion ranges between 55% and 60%, and this factor is also influenced by the spread of influence.

Inspired by the previous work which is [52], we empirically check whether there is a certain or similar pattern with target HI-index in terms of temporal effect. We found that target HI-index, indeed, did not perform well in a small dataset because the data sparsity issue is restricting this index from performing well. Through the random selection method that was repeated more than 1000 times, we can conclude that the newly designed time-awareness disparity seeding framework can reach fairness of gender ratio for seed selection, and we also found that the target HI-index also performs better compare to other indexes.

Research question 3. *Can we improve the disparity seeding algorithm by Utilizing temporal effect in social networks to achieve the goal of influence maximization?*

We conclude by answering that different types of diffusion process perform differently. We creatively came up with three different algorithms, origin seeding disperse algorithm (OSD), equal seeding disperse algorithm (ESD), Unequal seeding disperse Algorithm (USD). OSD was taken as our baseline algorithm. Considering the time factor. OSD has a relatively strong spread only in the early stage, and there is no significant growth in the middle and late stages. Because OSD lacks follow-up investment from seed users, it only invests a large number of seed users in the early stage. ESD has significantly increased the number and duration of users affected than OSD. We considered the influence of time factors in the diffusion stages, so we evenly distributed the seeds in different time intervals, which helped us gain the influence diffusion advantage over OSD. USD doesn't adopted to the equal seeding strategy. The spread of the distribution of dynamic allocation strategy is adopted in the process because we have adopted a more skilled allocation strategy. In the period of high user activity, we allocate more seeds, whereas, in the user active-low time, we assign fewer seeds to ensure that we have advantages in the overall dissemination.

We laid out the experimental results to answer the research questions in Section 5, OSD as our basedline algorithm. We found that USD perform better among the three algorithms. In particular, the optimal configuration which led to the best performance in our experiments is by choice the target ratio as close as the organic diffusion ratio. We also notice increased performance when considering sampling relative smaller number of users in relative smaller dataset. This shows the latent benefit that implicitly assume small scale social network communication scenarios.

6.2. Recommendation

Based on the experimental results and analysis performed in this thesis, some recommendations are proposed to our research group, which we hope might be useful to be considered in the future recommender system development at this research domain.

In the selection of seeds, we can do the targeted gender ratio filtering. However, we mainly rely on the target HI-index to adjust the neighbor's social influence control of the secondary transmission. However, I found that there might not be enough seed users to qualify because the graph size problem. When we consider the temporal effect, the whole time period cut into different time intervals. For example, in the first month, we want the target ratio to be 0.0, which means we filter out ten nodes that all affect women (which means their neighbors are all women), which is the perfect result. But in reality, there are not has such high number of users in each time slot as the static graph. Which means we do not have so many qualified nodes, resulting in the effect is not as good as the overall static graph. Therefore we proposed that we create a time-awareness target HI-index to target the temporal effect issue. The time-awareness target HI-index can based the size of the graph, dynamically adjust the requirements for the influence of neighbors and the gender ratio of neighbors' neighbors.

Since this experiment is based on the IC model, our seed users will not be activated once, which will affect the efficiency of communication in real communication. Because in many times the same seed user can be ranked at the top multiple times. Therefore, we suggest that we replace the underlying algorithm model for more extensive applications.

We also considered The Overlapping issue in this study. In the process of our research, we found that many top seed users may have similar interaction groups, which means that we may have many invalid seeds in the process of selecting seed users, and many seed users with high ranking can not really improve our communication efficiency too much. I believe that solving the Overlapping problem can greatly improve our algorithm results.

6.3. Future work

We acknowledge the lack of extensive evaluations of diffusion algorithms in our experiments. It is however intended because, given the time constraint, we focus on the evaluation framework rather than rigorous experiments comparing different algorithms. In the target in index in terms of the temporal effect scenario, more algorithms can be used to compare performance across different models, for example, machine learning algorithms and so on.

Different combinations of seed numbers are also key to our research. But also because of time and machine performance, we were not able to test more comprehensive seed combinations. If we were able to cover enough seed combinations, we might explore more conclusions and, therefore, more meaningful data. For example, seed 10000, 5000 and even more seeds are more meaningful in the INS dataset, but due to the time constraint of this experiment, we can not perform every seeding set in this experiment.

Furthermore, section 3 data analysis did not analyze the Target HI index data because we believe intensity and indegree are more representatives in measuring fundamental user interactions. If sufficient time and equipment are available in the future, a comprehensive statistical analysis of all indexes is a better choice.

References

[1] Charu C Aggarwal, Shuyang Lin, and Philip S Yu. “On influential node discovery in dynamic social networks”. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM. 2012, pp. 636–647.

[2] Junaid Ali et al. “On the fairness of time-critical influence maximization in social networks”. In: *arXiv preprint arXiv:1905.06618* (2019).

[3] Sergio Alonso et al. “h-Index: A review focused in its variants, computation and standardization for different scientific fields”. In: *Journal of informetrics* 3.4 (2009), pp. 273–289.

[4] Chen Avin et al. “Homophily and the glass ceiling effect in social networks”. In: *Proceedings of the 2015 conference on innovations in theoretical computer science*. 2015, pp. 41–50.

[5] Eytan Bakshy et al. “The role of social networks in information diffusion”. In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 519–528.

[6] Solon Barocas and Andrew D Selbst. “Big data’s disparate impact”. In: *Calif. L. Rev.* 104 (2016), p. 671.

[7] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems* 29 (2016), pp. 4349–4357.

[8] Matthew E Brashears. “A longitudinal analysis of gendered association patterns: Homophily and social distance in the general social survey”. In: *Journal of Social Structure* 16 (2015), p. 1.

[9] Matthew E Brashears. “Gender and homophily: Differences in male and female association in Blau space”. In: *Social Science Research* 37.2 (2008), pp. 400–415.

[10] Wei Chen, Wei Lu, and Ning Zhang. “Time-critical influence maximization in social networks with time-delayed diffusion process”. In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[11] Wei Chen, Yajun Wang, and Siyu Yang. “Efficient influence maximization in social networks”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 199–208.

[12] Alexandra Chouldechova. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *Big data* 5.2 (2017), pp. 153–163.

[13] Terry Connolly. *Micromotives and Macrobehavior.* 1979.

[14] IS Dhillon. “Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining”. In: (2001).

[15] Yuxiao Dong et al. “Structural diversity and homophily: A study across more than one hundred big networks”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 807–816.

[16] Golnoosh Farnad, Behrouz Babaki, and Michel Gendreau. “A unifying framework for fairness-aware influence maximization”. In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 714–722.

[17] Shanshan Feng et al. “Inf2vec: Latent representation model for social influence embedding”. In: *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE. 2018, pp. 941–952.

[18] Benjamin Fish et al. “Gaps in Information Access in Social Networks?” In: *The World Wide Web Conference*. 2019, pp. 480–490.

[19] Shay Gershtein et al. “IM balanced: influence maximization under balance constraints”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 1919–1922.

[20] Jacob Goldenberg, Barak Libai, and Eitan Muller. “Talk of the network: A complex systems look at the underlying process of word-of-mouth”. In: *Marketing letters* 12.3 (2001), pp. 211–223.

[21] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. “Celf++ optimizing the greedy algorithm for influence maximization in social networks”. In: *Proceedings of the 20th international conference companion on World wide web*. 2011, pp. 47–48.

[22] Mark Granovetter. “Threshold models of collective behavior”. In: *American journal of sociology* 83.6 (1978), pp. 1420–1443.

[23] Robert Y Justis and Barbara Kreigsmann. “The feasibility study as a tool for venture analysis”. In: *Journal of Small Business Management (pre-1986)* 17.000001 (1979), p. 35.

[24] Riittakerttu Kaltiala-Heino and Sari Fröjd. “Correlation between bullying and clinical depression in adolescent patients”. In: *Adolescent health, medicine and therapeutics* 2 (2011), p. 37.

[25] David Kempe, Jon Kleinberg, and Éva Tardos. “Influential nodes in a diffusion model for social networks”. In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2005, pp. 1127–1138.

[26] David Kempe, Jon Kleinberg, and Éva Tardos. “Maximizing the spread of influence through a social network”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 137–146.

[27] Moein Khajehnejad et al. “Adversarial Graph Embeddings for Fair Influence Maximization over Social Networks”. In: *arXiv preprint arXiv:2005.04074* (2020).

[28] Jinha Kim, Wonyeol Lee, and Hwanjo Yu. “CT-IC: Continuously activated and time-restricted independent cascade model for viral marketing”. In: *Knowledge-Based Systems* 62 (2014), pp. 57–68.

[29] Masahiro Kimura and Kazumi Saito. “Tractable models for information diffusion in social networks”. In: *European conference on principles of data mining and knowledge discovery*. Springer. 2006, pp. 259–271.

[30] Jure Leskovec et al. “Cost-effective outbreak detection in networks”. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, pp. 420–429.

[31] Kevin Lewis et al. “Tastes, ties, and time: A new social network dataset using Facebook. com”. In: *Social networks* 30.4 (2008), pp. 330–342.

[32] Kan Li, Lin Zhang, and Heyan Huang. “Social influence analysis: models, methods, and evaluation”. In: *Engineering* 4.1 (2018), pp. 40–46.

[33] Yuchen Li et al. “Influence maximization on social graphs: A survey”. In: *IEEE Transactions on Knowledge and Data Engineering* 30.10 (2018), pp. 1852–1872.

[34] Qiu Liqing et al. “Analysis of influence maximization in temporal social networks”. In: *IEEE Access* 7 (2019), pp. 42052–42062.

[35] Bo Liu et al. “Time constrained influence maximization in social networks”. In: *2012 IEEE 12th international conference on data mining*. IEEE. 2012, pp. 439–448.

[36] Vijay Mahajan, Eitan Muller, and Frank M Bass. “New product diffusion models in marketing: A review and directions for research”. In: *Journal of marketing* 54.1 (1990), pp. 1–26.

[37] K McDonald. *Physicist proposes new way to rank scientific output. Physorg, 2005*.

[38] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a feather: Homophily in social networks”. In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[39] Radosław Michalski, Jarosław Jankowski, and Piotr Bródka. “Effective Influence Spreading in Temporal Networks With Sequential Seeding”. In: *IEEE Access* 8 (2020), pp. 151208–151218.

[40] David S Moore and Stephane Kirkland. *The basic practice of statistics*. Vol. 2. WH Freeman New York, 2007.

[41] Shirin Nilizadeh et al. “Twitter’s glass ceiling: The effect of perceived gender on online visibility”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 10. 1. 2016.

[42] Naoto Ohsaka et al. “Dynamic influence analysis in evolving networks”. In: *Proceedings of the VLDB Endowment* 9.12 (2016), pp. 1077–1088.

[43] Travis E Oliphant. “Python for Scientific Computing, Computing in Science & Engineering”. In: (2007).

[44] Rohan Pearce. “Dead database walking: MySQL’s creator on why the future belongs to MariaDB”. In: *Computerworld, March* 28 (2013), p. 2013.

[45] Sancheng Peng, Guojun Wang, and Dongqing Xie. “Social influence analysis in social networking big data: Opportunities and challenges”. In: *IEEE network* 31.1 (2016), pp. 11–17.

[46] Fabıola SF Pereira et al. “On analyzing user preference dynamics with temporal social networks”. In: *Machine Learning* 107.11 (2018), pp. 1745–1773.

[47] Jiezhong Qiu et al. “Deepinf: Social influence prediction with deep learning”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2110–2119.

[48] Matthew Richardson and Pedro Domingos. “Mining knowledge-sharing sites for viral marketing”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 61–70.

[49] Andrea Saltelli. “Sensitivity analysis for importance assessment”. In: *Risk analysis* 22.3 (2002), pp. 579–590.

[50] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. “Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity”. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 923–932.

[51] B Stroustrup. “” Lecture: The essence of C++. University of Edinburgh”. In: *available from:(Accessed 12-Jun-2015) https://www. youtube. com/watch* 86 (2015).

[52] Arwen Teng et al. “On Influencing the Influential: Disparity Seeding”. In: *arXiv preprint arXiv:2011.08946* (2020).

[53] Jacqui True and Michael Mintrom. “Transnational networks and policy diffusion: The case of gender mainstreaming”. In: *International studies quarterly* 45.1 (2001), pp. 27–57.

[54] Quoc Dinh Truong, Quoc Bao Truong, and Taoufiq Dkaki. “Graph methods for social network analysis”. In: *International Conference on Nature of computation and Communication*. Springer. 2016, pp. 276–286.

[55] Petar Veličković et al. “Graph attention networks”. In: *arXiv preprint arXiv:1710.10903* (2017).

[56] Guojun Wang et al. “Fine-grained feature-based social influence evaluation in online social networks”. In: *IEEE Transactions on parallel and distributed systems* 25.9 (2013), pp. 2286–2296.

[57] Yu Wang et al. “Community-based greedy algorithm for mining top-k influential nodes in mobile social networks”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 1039–1048.

[58] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[59] Liying Wei. “Research on Google’s brand performance”. In: *Journal of Service Science Research* 8.2 (2016), pp. 161–175.

[60] Andreas Wimmer and Kevin Lewis. “Beyond and below racial homophily: ERG models of a friendship network documented on Facebook”. In: *American journal of sociology* 116.2 (2010), pp. 583–642.

[61] Stephan Winter, Caroline Brückner, and Nicole C Krämer. “They came, they liked, they commented: Social influence on Facebook news channels”. In: *Cyberpsychology, Behavior, and Social Networking* 18.8 (2015), pp. 431–436.

[62] Ke Yang and Julia Stoyanovich. “Measuring fairness in ranked outputs”. In: *Proceedings of the 29th international conference on scientific and statistical database management*. 2017, pp. 1–6.

[63] Jun Zhang et al. “Inferring continuous dynamic social influence and personal preference for temporal behavior prediction”. In: *Proceedings of the VLDB Endowment* 8.3 (2014), pp. 269–280.

[64] Honglei Zhuang et al. “Influence maximization in dynamic social networks”. In: *2013 IEEE 13th International Conference on Data Mining*. IEEE. 2013, pp. 1313–1318.