Evolving Information Cascading: Late Bird Matters

Zesen Zhang[1], Dongrui Lu[1], Luoyi Fu[1], Yingxiao Li[1], Xinbing Wang[1], Guihai Chen[1], Jun Xu[2]
Shanghai Jiao Tong University[1], China; Georgia Institute of Technology[2], USA.

ABSTRACT

Information cascading, also known as epidemic influence diffusion, is ubiquitous in social networks and vital for interpreting diverse social phenomena. To determine the (asymptotic) number of seed nodes required to spread a piece of information to virtually all nodes in a social network graph, through information cascading, is a well-studied research problem. However, most prior works study this problem only on a static graph (i.e., a snapshot of the social network). In this work, we study this problem in a more realistic setting: During the process of the information cascading, which could take a nontrivial amount of time (say months), the social network graph grows with the arrival of new nodes (newcomers) and their edges (social connections). This dynamic graph setting is a game changer since it brings about insights that cannot be learned in the static graph setting. For example, we find that newcomers command more influence power (ability to spread information) than existing nodes, since their social connections are "more interesting". We have studied this problem under two different graph topology (and growth) models, namely Erdos-Renyi (ER) and preferential attachment (RA), and three cascading models, and have obtained analytical results for all six combinations. Through these analytical results, we have gained considerable understanding of the collaborative impact of both models on the information cascading process. For example, ER growth model witnesses the 'reactivation' phenomenon and latter seeds could take the place of early seeds in PA growth model. Our theoretic claims are further validated by experiments in both simulated and real social networks.

1 INTRODUCTION

Ideas, technologies, diseases and choices can spread among people via social interactions, which in some cases can lead to a cascading behavior. Cascade is an epidemic process that begins with a set of seed nodes (i.e., seeds) that, by influencing other nodes and having the influenced nodes further propagate this influence, eventually leads to a massive outbreak across the network. This cascading process has been used to explain and/or to induce various phenomena in online social networks (OSN), such as viral marketing, smoking cessation and maintenance[1], micro-finance loans[2], and political campaigning via OSN[3]. Due to its practical importance, this cascading process has been a heavily studied research topic in the past few years.

An important research question along this topic is to determine, in terms of asymptotic order, the minimum number of seeds required to cause such an outbreak whose size is comparable to the whole network size. This research question has many real-life applications. For example, in OSN-based viral marketing, the marketer conceivably would like to influence, directly or indirectly, the vast majority of users in the OSN. While there have been considerable research studies on this research problem alone, most of them assume that the (OSN) graph is static during the information cascading process. This assumption is a bit detached from the reality, since the cascading process can take a nontrivial amount of time (say months), and in the meantime, the social network graph can grow considerably larger [4] with the arrival of new nodes (newcomers) and their edges (social connections). While readers might wonder whether analytical results under the static graph assumption can be suitably adapted (e.g., scaled by a growth factor) for dynamic graph scenarios, such adaptations are unlikely to work well, since they implicitly assume that existing users and newcomers have statistically identical influence powers, which we will show is not true.

In this work, we study information cascading in evolving (i.e., growing) social networks, focusing on the aforementioned research question of determining the minimum number of seeds required to induce an outbreak. Unlike in a static network where the seeds are determined upfront, in an evolving network seeds are usually selected on a continuous basis along with the network growth. For example, in viral marketing, often a new batch of seeds is resampled after each promotion period for the advertising to benefit from newcomers and their "refreshing" social connections, so that it can go beyond the old social circles of existing users to reach an ever larger population.

In this work, we choose seeds (uniformly) randomly, as opposed to most of the prior arts that try to find out the best seeds strategically, for the following reason. The random seed selection is often much more practically feasible than its strategic alternatives, such as that guided by the network structure information, since such information or knowledge can be extremely costly to acquire in field settings. Indeed, getting such information, such as the social connections between people in the network, is difficult, and influencing a specific (strategic) node/person can be very costly. For example, according to a recent estimation [5], conducting network surveys in 120 Indian villages would cost approximately $190, 000 and take over eight months. Furthermore, as shown in [6], in terms of cascading effectiveness, random seeding performs just as well as network-structure-guided seeding, and hence is strongly preferred in real-life situations with budgetary considerations.

We have found that the number of seeds necessary to cause an outbreak depends primarily on two factors: the information cascading model and the network topology. Both are indispensable: The former characterizes how a node is influenced by its neighbors and the latter to a large degree determines how far the information dissemination (i.e., influence spreading) process can go. In combination, they determine the scope and the rate of the diffusion process. In this work, we will consider all (six) combinations of three cascading models and two topology models.

A. Cascading Model: Generalizing k-complex contagion Models

The three cascading models we use in this work are the k-complex contagion model and two of its generalizations. In the k-complex contagion model, a node becomes infected when it is subject to the influence of at least k of its neighbors, where k is the influence threshold that is universal (i.e., for every node) across the network.

This model, widely used, is known to fairly accurately capture the phenomenon of multiple confirmation in many real-life scenarios, such as the adoption of expensive medical innovation, the decision to participate in a migration and changes in social behaviors[9][10]. For example, a study on Facebook showed that having two or more real-life friends already on Facebook substantially increases the probability of joining Facebook[11], and a similar phenomenon was observed in a study on Twitter [12]. For another example, statistics from DBLP and LiveJournal(LJ) [13] suggest that the second affected neighbor can often contribute more to influence spreading than the first.

A limitation of the k-complex contagion model is that the fixed threshold k does not capture the fact that individual nodes have different "risk tolerance levels": In real-life, some people are risk-taking (i.e., more willing to try out new things) while some others are risk averse. Hence, we generalize this model by making the threshold k of each node a random variable with a distribution $\mathscr{D}$. We use two different distributions in this work: uniform and Poisson. The former models the maximum possible diversity statistically in individual risk-tolerance levels whereas the latter has been empirically shown in [15] to fairly accurate capture the actual diversity in practice. These two generalized models, together with the original k-complex contagion model, are the three cascading models used in our analysis in the sequel.

B. Network Model: Evolving Topologies

As mentioned earlier, for our analysis (of the minimum seed size), we need also to assume a network topology (growth) model that specifies how the network graph grows with the arrivals of newcomers. In this work, we consider two such models: Erdös-Rényi(ER) graph [16] and Preferential Attachment model(PA model) [17]. In an ER graph, a newcomer connects to each of the existing nodes with a fixed probability p. Under this growth rule, latecomers are expected to have more initial edges. In a PA network, on the contrary, each node has the same number of initial edges. A new node connects to each existing node with a probability proportional to its degree. The resulting network has a power law degree distribution that is commonly seen in web graphs and academic networks. In the OSN context, ER model assumes indiscriminate selection of new acquaintances by each newcomer, whereas PA model assumes a preference toward those who have more acquaintances.

C. Our Results

We now specify the quantity we would like to analyze more precisely. For a target future graph size t (i.e., with t nodes at a future time), we want to know the minimum (asymptotic) number of seeds, uniformly randomly sampled from these t nodes over time (while the network is growing towards the target size t), that can influence the vast majority of the nodes in the network (i.e., cause an outbreak) before or when the network reaches the target size t. We perform this analysis under all aforementioned six combinations of network topology models and cascading models, in an effort to obtain a comprehensive understanding of the relation between this minimum seed size and the target network size t.

A key observation we can make from these analyses is that, in nearly all six cases, newcomers (i.e., late birds) are generally more important in influence spreading. Beside that, other surprising results are summarized as follows:

(1) In ER graphs, when the influence threshold k is a random variable, the required seed value decreases with the target network size t and approaches a constant value when t tends to infinity. We will show this counter-intuitive finding confirms the aforementioned insights that latecomers have more influence power.

(2) In ER graphs, with a nonzero probability, the diffusion process may get stuck, in the sense no more nodes can be further influenced, before or when it grows to the target size t. Under such circumstances, the process can be "rejuvenated" by recruiting a sufficiently large number of additional nodes into the network (beyond the target size t), even when none of these additional nodes is a seed. This phenomenon further demonstrates the disproportional influence power of the late birds.

(3) In PA networks, the required minimum seed size grows with the target network size t, so in a sense the PA model is less favorable to the information diffusion process. In the worst case, this seed size can $O(\sqrt{t})$. However, in this case, in terms of the percentage of target network size t, this seed size actually decreases when t is larger, which again confirms the information diffusion power of late birds.

Our analyses of the minimum required seed size are further confirmed by experiments under simulated and real networks. Due to space limitations, many detailed proofs and derivations involved are omitted and placed in our full version [31]. To our best knowledge, this is the first effort made towards understanding the information cascading in temporally growing networks. By disclosing asymptotically how seed size governs the cascading during network evolution, we believe our results are of fundamental importance for better understanding the dynamics of epidemics in real systems.

2 RELATED LITERATURE

In this section, we give a brief review of the related researches on information diffusion. Generally speaking, efforts have been made in two aspects: cascading and influence maximization. Since both are classic research fields with fruitful achievements, we simply list the prior works closest to ours.

Cascading focuses on revealing the diffusion schema in certain surroundings. Diffusion in PA networks under IC or LT model has been extensively studied[18][19]. When the initial seeds are chosen uniformly at random, the diffusion is also called the bootstrap percolation[20]. Bootstrap percolation is examined in various networks, including ER graphs[21]. [22] studies the problem of finding a set whose infection would trigger an influence scale comparable to the entire population in a static network. [15] and [23] analyse the cascade in time-evolving networks under k-complex contagion model and they are the most relevant to our work. Yet their conclusions are drawn from rather different perspectives.

Influence maximization (IM), however, concentrates on algorithm design. Influence maximization aims to maximize the influenced members by determining an optimal seed set with a fixed size. First introduced by Kempe et al., IM has been studied in countless works from multiple angles. Most algorithms are designed for IM in static contexts. Recent researches have focused on proposing improvements and modifications under various constraints. For example, [7] and [8] address the algorithm scalability and propose solutions for billion-scale networks. [14] analyses IM under discount constraints. Meanwhile, there emerges a new class of IM where seeds are chosen periodically in dynamic networks, which is most similar to our settings. [24] studies IM in networks with changing edges and fixed nodes. [25] and [26] periodically select seeds according to influence estimation based on previous diffusion feedbacks via a bandit-based approach.

To conclude, there have been vast studies both in cascading and IM, but to the best of our knowledge, no attempts have been made in determining the average number of seeds required to expand influence to the whole network in a dynamic context.

3 PRELIMINARIES

3.1 Cascading Model

We embrace the k-complex contagion model to describe the influence diffusion in undirected graphs. We would like to first introduce several notions of node which we will use in later sections, and then the definition of the generalized k-complex contagion model.

Definition 3.1. ('Blank' Node, 'Influenced' Node, 'Influenced' Degree, Seed): At a given time stamp, 'Blank' nodes refer to nodes who have not yet been influenced and 'influenced' nodes are nodes that have already been influenced. A 'blank' node can be turned into an 'influenced' node either during initialization or later by its neighbors. And we call the degree that an 'influenced' node has as an 'influenced' degree. For example, if there are ten edges connect with an 'influenced' node, than we say the network has ten 'influenced' degrees. Seeds are nodes who are influenced during initialization. They are the source of the information cascading.

Definition 3.2. (K-Complex Contagion Model): Given an undirected graph G generated by an evolving process, a k-complex contagion $CC(G, p_0, R_u, \mathscr{D})$ is a contagion initiated by a p_0 proportion of nodes (seed proportion being p_0 during initialization) that spreads across network G. The information diffusion proceeds in rounds. R_u is a stochastic variable that follows distribution $\mathscr{D}$. It represents the node 'influence threshold' k. The 'influence threshold' k of each node is generated in round 0 and is fixed during information diffusion. By the end of each round, every node with no less than k influenced neighbors is influenced.

K-complex contagion model is a general model that captures the influence spreading dynamics in a network. Let us take the popularization of smart phones as an example to illustrate the process. When smart phone was first introduced, people who were using traditional phones were 'blank' nodes and those who adopted smart phones in the first place were seeds. Assume a person who was still using traditional phone. He was a 'blank' node. One day, he found one of his friends using smart phone and heard positive remarks on the smart phone from this friend. Several weeks later, he found a great number of his friends starting to use smart phones and he heard a lot more good positive evaluations. He was finally influenced by his friends and decided to buy a smart phone. In doing so, he became an 'influenced' node. Further, if he found smart phone convenient and recommended it to others, he would be contributing to the popularization of the smart phone, aka. the information diffusion. In this example, the influence threshold k is the number of friends who use smart phones the moment a person decides to switch to a smart phone.

In this paper, we go beyond the tradition k-complex contagion model which simply assumes that each node has the same 'influence threshold' k. The K-complex contagion model we adopt conforms more to the reality in that k varies from individual to individual: some people tend to go with stream, while others are likely to stick to their own idea. We consider two different distributions $\mathscr{D}$ of the 'influence threshold': *Uniform Distribution* and *Poisson Distribution.* The term $R_u \sim \mathscr{D}$ specifies that 'influence threshold' k follows distribution $\mathscr{D}$.

3.2 Graph Model

An evolving general network is a network that has an identical evolution dynamics of a real network. The following theorem shows the difficulty of our problem in an evolving network with general topology.

THEOREM 3.3. *Finding the minimum number of seeds to cascade the influence to the whole evolving general network is an NP-hard problem.*

The main idea of the proof is to reduce this problem to the bond percolation one, which can be further converted to the Set Cover Problem. The specific proofs of Theorem 3.3 can be found in full version [31]. Given the NP-hardness of the problem, we study instead the information diffusion under two specific evolving network models, i.e., the evolving Erdös-Rényi (ER) graph and the evolving Preferential Attachment (PA) network.

An ER network, or alternatively denoted as $G(n, p)$ graph, has been widely served in a large body of literature as the basic model due to its enjoyable mathematical tractability. In a $G(n, p)$ model, there are n nodes in total and an edge exists between any node pair with probability p independently. p can also be interpreted as graph density, as an increase in p yields a denser graph. Based on the static ER model, we define below its evolving version.

Definition 3.4. (Evolving Erdös-Rényi Model): In an *Evolving Erdös-Rényi model* (evolving ER Model, network grows at a uniform rate, with one new node being added to the network at each time slot. Each new node is supposed to emit an edge to every existing node with the same probability p. In other words, each new node is expected to emit a total of tp edges upon arrival, with t being the network size the moment the new node is added. The evolving ER Model incorporates the growth nature of social networks. As p stays fixed and the network continues to expand, latter nodes will have more initial edges than the earlier ones.

While ER graphs can also reflect the equal chance of connection among different users, in some other real situations new comers have tendency to connect to those of higher popularity. Take Twitter for instance, users are more inclined to follow the account of an influential figure like Taylor Swift or a big organization like China Daily. The same also holds in academic networks, where papers that have higher citations have a higher chance of being further cited. This common biased connection choice can be well depicted by the celebrated Preferential Attachment Model. In the evolving networks of interests, we formally define the Evolving Preferential Attachment Model as follows.

Definition 3.5. (Evolving Preferential Attachment Model): In *the Evolving Preferential Attachment Model* (the evolving PA model) $PA_{m,n}(V, E)$, new nodes come into the network in a sequence. One new node comes at every time slot and emits m edges to the the existing nodes. The probability that an existing node v connects to the latest node is proportional to its degree $deg(v)$ (the preferential attachment rule). Note that $\sum_{v \in V} deg(v) = 2mn$. Therefore, these nodes are supposed to gain more degrees than nodes that come later. The generated network has a power law degree distribution.

As noted earlier, when considering the diffusion process in two above network models, we adopt random seeding that goes in parallel with network evolution. Meanwhile, the 'influence threshold' k of each node is determined upon arrival and stays fixed. Thus, for a target future graph size t (i.e., with t nodes at a future time), we aim to derive the minimum (asymptotic) number of seeds, uniformly randomly sampled from these t nodes over time (while the network is growing towards the target size t), that can influence the vast majority of the nodes in the network before or when the network reaches the target size t [1].

4 MAIN RESULTS

We present in this section the main results of this paper with the above definitions and statements along with their intuitive interpretations. The proofs are relegated to Sections 5. We are going to unfold the estimated order of seed quantity required to achieve an influence scale comparable to or bigger than the whole network under six settings. Each setting is a combination of a network topology (an ER graph or a PA graph) and a diffusion model (one of the three k-complex contagion models).

4.1 Larger Network Size But Fewer Seeds

Before unfolding the main results in ER networks, we would like to recall that the 'influence threshold' R_u always equals k_0 in traditional k-complex contagion model, but follows the distribution $\mathscr{D}$ in generalized k-complex contagion model. When $\mathscr{D}$ stands for *Uniform Distribution*, $R_u \in [\![1, n-1]\!]$. When $\mathscr{D}$ refers to *Poisson Distribution*, we normalize $\sum_{l=1}^{n-1} e^{-\lambda} \frac{\lambda^l}{l!}$ to 1. With these preliminaries, we are able to draw conclusions when $n \longrightarrow \infty$.

Theorem 4.1. *Let $G(n,p)$ and $CC(G, p_0, R_u, \mathscr{D})$ represent respectively the evolving ER Model and the k-complex contagion model. In order to achieve an influence scope that has the same order of the whole network till time slot n, the probability that each new node becomes a seed should satisfy*

(1) $\Theta\left(\frac{\ln\left(\frac{\sqrt{k_0}n}{n-1}\right)}{np}\right)$ *under traditional k-complex contagion model ($k = k_0$).*

(2) $\frac{\Theta((1-p))}{1-\Theta(p)}$ *under generalized k-complex contagion model where k follows the uniform distribution.*

(3) $\Theta\left(\frac{\ln\left(\frac{\sqrt{\lambda}n}{n-1}\right)}{np}\right)$ *under generalized k-complex contagion model where k follows the Poisson distribution.*

Theorem 4.1(2) shows that under generalized k-complex contagion model where $\mathscr{D}$ obeys to a uniform distribution, the seed quantity increases along with the network expansion. Moreover, it is comparable to n. Since $R_u \in [\![1, n-1]\!]$, the uniform distribution augments average "influence threshold" during network evolution, which intuitively impedes the influence from spreading.

In contrast, **Theorem 4.1**(1) and (3) imply that when k is fixed or follows a Poisson distribution, fewer seeds are needed for a network-wide influence scale as network expands. In fact, an order of $\Theta\left(\frac{1}{p}\ln\left(\frac{\sqrt{k_0}n}{n-1}\right)\right)$ ($\Theta\left(\frac{1}{p}\ln\left(\frac{\sqrt{\lambda}n}{n-1}\right)\right)$) seeds are sufficient to trigger an influence range comparable to the entire network. This seemingly counter-intuitive phenomenon suggests strong diffusion power of late birds. Recall that in the evolving ER model, late birds are supposed to emit more edges to the network. Higher degrees of late birds contribute to their strong cascading capability, which, in some cases, can lead to a 'reactivation' phenomenon.

Definition 4.2. ('Reactivation' Phenomenon): The 'reactivation' phenomenon is a phenomenon that sometimes occur in ER networks under the k-complex contagion model. For simplicity, we adopt the 2-complex contagion model and illustrate the situation in figure 1. The diffusion is at an impasse at some time slot (the leftmost picture). In the next time slot, a new node connects to the only two infected nodes and two other uninfected nodes. Thanks to the newcomer, the diffusion is 'reactivated' and eventually the whole network is influenced.

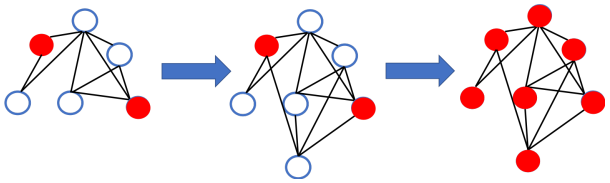


Figure 1: 'Reactivation' phenomenon. Late birds could recur the influence diffusion.

In order to consider 'reactivation' in ER networks, we define the Advanced Evolving ER model $G(n, n', p)$ based on 'reactivation'.

Definition 4.3. (Advanced Evolving ER Model): In $G(n, n', p)$, we divide network evolution into two phases, with network size being n and n' by the end of each phase. During the first stage, we will randomly pick some nodes to be seeds. The cascading reaches an impasse by the end of the first phase. In the second stage, we add another $n' - n$ 'blank' nodes into the network, hoping to 'reactivate' the diffusion process.

In the following sections, we call the first phase as the 'impasse' part and the second phase as the 'reactivation' part.

The advanced evolving ER model $G(n, n', p)$ states that when the diffusion process stagnates by the end of 'impasse' part, the arrival of 'blank' newcomers in 'reactivation' part can help 'revive' the influence spreading under certain conditions. We hereafter would like to precise the conditions under each diffusion model in the following lemma.

Lemma 4.4. *Let p_0 be the proportion of influenced nodes by the end of the first phase of network evolution. Nodes added during the second phase can reactivate and boost the influence scope comparable to network size till time 'n' if the following conditions are met:*

(1) Under traditional k-complex contagion model where k is fixed, the network-wide influence diffusion happens with probability $1 - \Theta(\frac{1}{n})$ if $np_0 \geq k_0$ and n' satisfies

$$n' \geq \Theta\left(\frac{k_0 + 2\ln n - np_0 p}{(n-1)(n-2)p^2 p_0^2} + n\right) \tag{1}$$

(2) Under generalized k-complex contagion model where k follows the Uniform Distribution, when $R_u \in [\![1, n-1]\!]$, the network-wide diffusion happens with probability $1 - \Theta(\frac{1}{n})$ if n' satisfies

[1] In the rest of the paper, we will use t and n interchangeably.

$$n' \geq \Theta\left(\frac{n}{p_0 p}\right) \quad (2)$$

(3) Under generalized k-complex contagion model where k follows Poisson distribution and $R_u \in [\![1, n-1]\!]$, the network-wide diffusion occurs with probability $1 - \Theta(\frac{1}{n})$ if n' satisfies

$$n' \geq \Theta\left(\frac{2\ln n - np_0 p + \lambda e^{t_1}}{\sqrt{\lambda}}(n-1) + n\right) \quad (3)$$

where t_1 is a constant.

Lemma 4.4 elaborates the claim that late birds can reactivate the influence diffusion in ER networks when n' satisfies a certain inequality. A closer examination of the inequality constraints leads to the conclusion that the 'reactivation' is most unlikely to occur when k follows the *uniform distribution.* Furthermore, the 'reactivation' is more difficult to take place under the generalized k-complex contagion model than under traditional k-complex contagion model. Mathematically, the inequality (2) is the most strict condition mainly because the uniform distribution has highest expected number of k, which impede information diffusion mostly. And the inequality (3) is harder than inequality (1) due to the fact that nodes with high k under generalized k-complex contagion models impede greatly in the cascading. The 'reactivation' process also explains the decline of the required proportion of seeds in some cases in **Theorem 4.1** during network expansion.

4.2 Late Comers but Large Diffusion Power

Next, we would like to show our results in PA networks $PA_{m,n}(V, E)$. Similarly, in generalized k-complex contagion models, we will let $R_u \in [\![1, m]\!]$ when the distribution $\mathscr{D}$ stands for *Uniform Distribution.* Otherwise, we will normalize the expression $\sum_{l=1}^{m} e^{-\lambda} \frac{\lambda^l}{l!}$ to 1 when $\mathscr{D}$ represents *Poisson Distribution.* Then we are able to draw conclusions even when $n \longrightarrow \infty$.

Theorem 4.5. *Denote the Evolving PA Model to be $PA_{m,n}(V, E)$ and the k-complex contagion model to be $CC(G, p_0, R_u, \mathscr{D})$. In order to achieve an influence scope that covers the whole network, the probability of nodes to become seeds upon arrival during the first phase should satisfy:*

(1) $\Theta\left(\frac{1}{\sqrt{n}-\ln(n-1)}\right)$ under traditional k-complex contagion model where k is fixed.

(2) $\Theta\left(\frac{1}{\sqrt{n}(n+H)^H}\right)$ under generalized k-complex contagion model where k follows the uniform distribution and $H = \frac{(\sqrt{m-1}-1)}{m^2\sqrt{2\pi}}$.

(3) $\Theta\left(\frac{1}{\sqrt{n}-\ln(n-1)}\right)$ under generalized k-complex contagion model where k follows Poisson distribution.

Theorem 4.5 shows that as PA networks grow larger, the necessary seed quantity also increases. This suggests that the PA based network structure somewhat prevents influence from spreading. The reason may be that the 'reactivation' phenomenon seldom occurs. In PA networks, the 'reactivation' occurs with a probability of p^S, where S is a constant in a specific k-complex contagion model. In the evolving PA model, as [27] shows, the highest probability that the new node connects to an existing node is $\frac{m}{\sqrt{n}}$, with n being the network size. Therefore, 'reactivation' occurs with a probability smaller than $\left(\frac{m}{\sqrt{n}}\right)^S$. As $n \longrightarrow \infty$, 'reactivation' is increasingly unlikely to occur. Nonetheless, we can observe that the seed proportion decreases as PA networks evolve, which again shows the diffusion capability of late birds.

Moreover, as we could see from the **Theorem 4.5**, under the traditional k-complex contagion model, the order of seed quantity is $\Theta(\frac{n}{\sqrt{n}-\ln(n-1)})$. When a network doubles its size, the order of seed quantity becomes $\Theta(\frac{2n}{\sqrt{2n}-\ln(2n-1)})$, about $\sqrt{2}$ times the former quantity. Because the seeds distribute uniformly in the network, this relation suggests that some late birds (about $\sqrt{2}+1$) can exert approximately the same amount of influence on cascading as one early bird. Considering the fact that late coming nodes in PA networks have smaller degree in the network, the scaling relation is a meaningful discovery and can have practical interests in reality. Take advertising campaign for example, asking a celebrity for endorsement costs much more than randomly asking a dozen of people for online product promotion.

5 PROOF OF THE RESULTS

In this section, we prove that with a certain number of randomly-assigned seeds coming, influence can eventually be spread to the scale comparable to the whole evolving ER (or PA) network size under generalized k-complex contagion models when the network grows big enough.

5.1 Overview of The Proof

We would like to provide a proof overview before diving into technical details. Our proof is exclusively for undirected contagions, where influences can diffuse in both directions on an edge. As seeds come randomly into the network along with 'blank' nodes, we can assume seeds come at a probability of θ. And our goal is to find a suitable θ to spread the influence to the scale comparable to whole network size with high probability.

The diffusion process can be decomposed into two parts: 'initial' diffusion and 'inner' diffusion. The 'initial' diffusion shows the influence that newly coming nodes bring to the network. During 'initial' diffusion process, the newly coming node is a seed or it get influenced by other nodes in the network, it may successfully influence its neighbor nodes who already exit in the network. During 'inner' diffusion, influence spread between existing nodes. Newly influenced nodes in the network spread influence to their neighbors in an iterative manner. In the following we present the analysis of "initial diffusion" and "inner diffusion" in mathematical ways.

'Initial' Diffusion: Let G be the generated graph according to the evolving ER or PA model. Assume that node t is the t_{th} node that arrives in G. Let V_t be the set of first t nodes in G and X_t be the set of 'influenced' nodes in V_t. If t has an 'influence threshold' k, t gets infected if and only if at least k of its neighbors belong to X_{t-1}. Let I_t be the state indicator of node t, with $I_t = 1$ signifying the infected state and $I_t = 0$ otherwise. We use $\mathbb{1}(N_u^t) = 1$ to symbolize the event that node u connects with node t. Otherwise not. Therefore, $\mathbb{1}(N_u^t) \cdot I_t = 1$ means the node u has an influenced number t. Define Y_t as the current proportion of the 'influenced' degrees (recall Definition 3.1) in the whole network. Note the degree of node t as $deg(t)$. Then, $\sum_{v \in V_t} deg(v)Y_t$, the total number of the 'influenced' degrees in the whole network after the arrival of t, is determined by the following four ingredients:

- There are currently $\sum_{v \in V_{t-1}} deg(v)Y_{t-1}$ 'influenced' degrees.
- The edges that t emit into the set X_{t-1} will contribute to the infected degrees.
- Let p_{sum} be the probability that node t gets infected. p_{sum} is a sum of two probabilities: the probability that t is a seed and the probability that t is immediately influenced upon arrival. If t is infected, then all the edges it emits contribute to the 'influenced' degrees.
- Let $H(u)$ be the number of infected neighbors of an existing node u whose 'influence threshold' is k. Suppose now $H(u) = k - 1$. Thus, u will become influenced if t is influenced and t connects to it. We multiply the probability of this event as $M(u) = \Pr[I_u | N_u^t, H(u) = k-1, I_t = 1, I_u = 0]$. Multiplying $M(u)$ with the degree of u helps us get the expected number of 'influenced' degrees node u could bring to the network under the above phenomenon. Then we use γ to denote the whole 'influenced' degrees the above event could bring to the network.

Adding up these ingredients, we get the following recurrence:

$$\sum_{v \in V_t} deg(v)Y_t = \sum_{v \in V_{t-1}} deg(v)Y_{t-1} + deg(t)Y_{t-1} + deg(t)p_{sum} + \gamma, \tag{4}$$

$$p_{sum} = \theta + (1-\theta)\Pr[I_t = 1], \tag{5}$$

$$\gamma = \sum_{u=1}^{t-1} \left(\Pr[I_u | N_u^t, H(u) = k-1, I_t = 1, I_u = 0] \cdot deg(u) \right). \tag{6}$$

'Inner' Diffusion We assume that the influence diffusion is instantaneous. Therefore, we do not have to consider actual start point of the 'inner' diffusion and can directly depict the final state of 'inner' diffusion at last time slot. In evolving ER networks, as all nodes have the same probability p to connect to others, the diffusion pattern of each node is identical. Therefore, we could analyze the 'inner' diffusion via a holistic approach. In evolving PA networks, however, we always observe a cumulated 'inner' diffusion. As the influence diffuses without delay, we can release the 'inner' diffusion at the last time slot.

According to the k-complex contagion model, a 'blank' node will not become an 'influenced' node unless it is initially influenced in the 'initial' diffusion or surrounded by no less than k 'influenced' neighbors, with k its 'influence threshold'. Therefore, let p_0 be the proportion of influenced nodes by the end of the 'initial' diffusion, and i, u be the node in network, we have:

$$\Pr[I_u = 1] = p_0 + (1-p_0) \left(\sum_k \Pr_{R_u \sim \mathscr{D}}[R_u = k] \cdot \Pr[\sum_{i=1}^{n} \mathbb{1}\left(N_u^i\right) \cdot I_i \geq k] \right) \tag{7}$$

Then we can get the probability that the scale comparable to the whole network size is affected by multiplying the infection probability of each node u.

Therefore, by combining the 'initial' and 'inner' diffusion process together, we can deduce the suitable θ that ensures a network-wide infection.

To make our proof easier to demonstrate, we first show the power of the 'inner' diffusion and find the suitable proportion of influenced nodes to spread the influence to the whole network. Then, with recurrence, we consider the 'initial' influence and inversely derive the number of seeds.

5.2 Cascading under Traditional k-complex contagion Model

In this part, we will first prove our results in evolving ER networks and then in evolving PA networks. Due to space limitations, we only unfold our most important technical flow in the proof. Many intermediate results are summarized in the form of lemmas and corollaries, whose detailed derivations can be deferred to the full version [31].

5.2.1 Proof of Lemma 4.4 (1). This lemma shows the power of nodes that arrive in the 'reactivation' part (recall the Definition 4.3) in $G(n, n', p)$ and reveals the 'inner' influence among nodes of the 'impasse' part (i.e. the 'reactivation' phenomenon). We assume i as the nodes in the 'reactivation' part and u as the 'blank' node in the 'impasse' part. As Fig.1 shows, in order to contribute to the influence diffusion, the node i in 'reactivation' part needs first to be influenced (i.e. $I_i = 1$) and then to connect to the 'blank' nodes u (i.e. $N_u^i = 1$) in the 'impasse' part. In order to be influenced, a node i from the 'reactivation' part needs to emit k edges into the 'influenced set' X_n (recall the definition in the 'initial' diffusion). And the probability that i emits an edge into X_n is pp_0. Therefore, we use bayes formula to denote the probability that a node in 'reactivation' part facilitates influence spreading in the following way:

$$\begin{aligned} \Pr[I_i | N_u^i, I_u] &= \frac{\Pr[I_i = 1, N_u^i, I_u = 0]}{\Pr[N_u^i, I_u = 0]} \\ &= \frac{\sum_{u=k}^{n} C_n^u \, (p_0 p)^u \, (1 - p_0 p)^{n-u} \, (1 - (1-q)^{n-u})}{1 - (1 - (1 - p_0 p)q)^n} \end{aligned} \tag{8}$$

Here the 'q' in the above expression refers to the inverse of the 'uninfected' population in the 'impasse' part.

Suppose the fixed 'influence threshold' R_u is k. In an evolving ER network, every edge emerges with probability p. Recall that the proportion of influenced nodes is p_0 by the end of the 'impasse' part. Consider a node u in in $G(n, n', p)$. The influence that u is submitted to consists of two parts: the influence originating from nodes in the 'impasse' part and that from the 'reactivation' part. Nodes from the 'impasse' part impose a total influence of npp_0 to u. Nodes from the 'reactivation' part give u an influence that amounts to $\sum_{i=n+1}^{n'} \Pr[I_i | N_u^i, I_u]$. According to Eqn. (7), the probability that u gets influenced in an evolving ER network is:

$$\begin{aligned} \Pr[I_u = 1] &= p_0 + (1-p_0) \left(\Pr[\sum_{i=1}^{n'} \mathbb{1}(N_u^i) \cdot I_i \geq k] \right) \\ &= p_0 + (1-p_0)\Pr[p_0 pn + \sum_{i=n+1}^{n'} \Pr[I_i | N_u^i, I_u] \geq k] \end{aligned} \tag{9}$$

Then we can further simplify $\Pr[I_i | N_u^i, I_u]$ with the following corollary.

Corollary 5.1. *When $n' \to \infty$ and $p << 1$, we have:*

$$Pr[I_i | N_u^i, I_u] = (n-1)(n-2)(p_0 p)^2 \tag{10}$$

Injecting the corollary into the expression of $\Pr[I_u = 1]$, we have:

$$\Pr[I_u = 1] = p_0 + (1-p_0) \left(1 - \Pr[p_0 pn + (n'-n)(n-1)(n-2)p_0^2 p^2 < k] \right)$$

Hence, the probability that every node in the 'impasse' part is influenced is:

$$\begin{aligned} &\Pr\left[\text{First } n \text{ number of nodes get influenced}\right] \\ &= 1 - n(1-p_0)\Pr\left[\sum_{i=1}^{n'} \mathbb{1}\left(N_u^i\right) \cdot I_i < k\right] \end{aligned} \tag{11}$$

Assuming t_1 as a constant, by *chernoff bound* we have:

$$\begin{aligned} &\Pr[\text{First } n \text{ number of nodes get influenced}] \\ &= 1 - \frac{n(1-p_0)\exp(t_1 k)}{\exp\left(t_1 \cdot \left(p_0 p n + (n'-n)(n-1)(n-2)p_0^2 p^2\right)\right)} \end{aligned}$$

Once $np_0 p + (n'-n)(n-1)(n-2)p_0^2 p^2 \geq k + 2\ln n$ is satisfied, Pr[First n number of nodes get influenced] will equal to $1 - \Theta(\frac{1}{n})$, which goes to 1 when $n \to \infty$.

Furthermore, Lemma 4.4 (1) shows that once $n' = n$ (i.e. when all the nodes belong to the first part), the 'inner' influence booms up when $np_0 p = \Theta(2\ln n)$. With this result, we proceed to give out the proof of Theorem 4.1 (1).

5.2.2 *Proof of Theorem 4.1(1).*

The key to the proof is to explicit the 'influence' potential (i.e. p_{sum} and γ) of the newcomer node. As new nodes come into the network iteratively, influence spreads recursively. Therefore, we try to establish recursive equations and find the expressions of p_{sum} and γ.

p_{sum} is affected by the following two probabilities:

- The probability that the newly coming node is a seed.
- The probability that the newly coming node connects with a sufficient number of 'influenced' nodes.

As such, we can rewrite Eqn. (5) for an elaborated p_{sum}:

$$p_{sum} = \theta + \left(1 - \sum_{u=0}^{k-1} C_{t-1}^{u} \left(pY_{t-1}\right)^{u} \left(1 - pY_{t-1}\right)^{t-u-1}\right)(1-\theta) \tag{12}$$

Similarly, γ can be elaborated by rewriting Eqn. (6) as:

$$\gamma = (t-1)^2 p^2 p_{sum} C_{t-1}^{k-1} (pY_{t-1})^{k-1} (1 - pY_{t-1})^{t-k} \tag{13}$$

Both p_{sum} and γ can be further simplified by the following corollary.

Corollary 5.2. *When $k = \Theta(1)$ and $t \to \infty$, p_{sum} and γ could be simplified as $p_{sum} \approx \frac{2\sqrt{k}}{\sqrt{2\pi}(t-1)}$ and $\gamma \approx (t-1)p^2 p_{sum} \frac{\sqrt{t-1}}{\sqrt{2\pi}\sqrt{(k-1)(t-k)}}$.*

In an ER network, all nodes share the same degree expectation: $deg(t) = (t-1)p$ and thus possess same 'status'. By sequentially introducing new nodes and bringing their 'influence' power into the network, we can rewrite Eqn. (4) as:

$$\begin{aligned} tY_t &= (t-1)Y_{t-1} + \\ &\left(\left(1 - \frac{2\sqrt{k}}{\sqrt{2\pi}(t-1)}\right)\theta + \frac{2\sqrt{k}}{\sqrt{2\pi}(t-1)}\right)\left(1 + \frac{p}{\sqrt{2\pi(k-1)}}\frac{\sqrt{t-1}}{\sqrt{t-k}}\right) \end{aligned} \tag{14}$$

The following corollary concerns Y_n.

Corollary 5.3. *As $n \to \infty$ and $n >> k$ we can reform the equation with several constants represented as capitalized letters:*

$$nY_n = (An - B\ln(n-1) + C)\,\theta + D\left(\ln \frac{\sqrt{2\pi}}{2\sqrt{k}}(n-1)\right) - E \tag{15}$$

Here, $A = 1 + \frac{p}{\sqrt{2\pi(k-1)}}$, $B = \frac{2\sqrt{k}}{\sqrt{2\pi}} + \frac{p}{\pi} - \frac{p \cdot \sqrt{k-1}}{2\sqrt{2\pi}}$, $C = \frac{\sqrt{k-1}p}{2\sqrt{2\pi}} \ln 2 - \frac{\sqrt{k-1}pln(\sqrt{k-1})}{\sqrt{2\pi}} - \frac{2\sqrt{k}}{\sqrt{2\pi}} \ln(\frac{\sqrt{2\pi}}{2\sqrt{k}}(k-1)) - \frac{p}{\pi}\ln 4 - \frac{2\sqrt{k}p}{\sqrt{2\pi(k-1)}}\ln(k-1)$, $D = \frac{2\sqrt{k}}{\sqrt{2\pi}} + \frac{\sqrt{2}p}{\sqrt{\pi}}$, $E = \frac{2p}{\sqrt{2\pi}}\ln(k-1) + \frac{\sqrt{2k}}{\sqrt{2\pi}}\ln\left(\frac{\sqrt{\pi}}{\sqrt{2k}}\right)(k-1)$.

The above proof shows that once $npp_0 = \Theta(2\ln n)$, the 'inner' diffusion is powerful enough to spread the influence to the whole network. Therefore, we have $nY_n = \Theta(\frac{2\ln n}{p})$. To make our expression neat, we only keep the dominating term in the expression. Hence, we can draw the conclusion that in ER networks under traditional k-complex contagion model, the network wide influence spreading occurs if θ satisfies $\theta = \Theta\left(\frac{\ln\left(\frac{\sqrt{k}n}{n-1}\right)}{np}\right)$.

5.2.3 *Proof of Theorem 4.5(1).*

As every node in an evolving PA network emits m edges upon arrival, with $m = \Theta(1)$, the 'reactivation' has little chance to take place. In view of this, we neglect $\Pr[I_i | N_u^i, I_u]$ in PA networks.

Our proof is built upon two corollaries. Corollary 5.4 explicates the probability for a node to get influenced and the proportion of influenced nodes required for a network-wide influence scale. Corollary 5.4 analyzes the cases where the 'inner' diffusion can produce a network-wide cascading. Corollary 5.5 shows the potential 'influence' power of the newly coming node (i.e. 'initial' diffusion). Particularly, we show that the scale that comparable to whole network size will get influenced with probability $1 - \Theta(\frac{1}{n})$ when the initial seed proportion θ meets certain condition.

Corollary 5.4. *Suppose $p_0 > 0$ and let $d(u)$ be the degree of node u. Then we specify Eqn (7)*

$$Pr[I_u = 1] = p_0 + (1-p_0)\left(1 - Pr\left[\sum_{i=1}^{n} \mathbb{1}\left(N_u^i\right) \cdot I_i < k\right]\right) \tag{16}$$

where

$$Pr\left[\sum_{i=1}^{n} \mathbb{1}\left(N_u^i\right) \cdot I_i \leq k\right] = Pr\left[d(u)\left(p_0 + \frac{\sum_{j=1}^{u-1} d(j)}{2mn}(1-p_0)\right) < k\right].$$

Furthermore, if $p_0 = \Theta\left(\frac{\sqrt{n}}{n}\right)$, the cascading scope that is comparable to the whole network size will occur with high probability.

Corollary 5.5. *Let k and m be two constants with $m > k$. Let $L = \frac{k}{2m} + \frac{3}{2} + \frac{m\ln\frac{m-k}{m}}{2k}$, $M = \frac{1}{\sqrt{2\pi(k-1)}}$, $N = \frac{2\sqrt{k}}{\sqrt{2\pi}(m-1)}$ be constants. Under evolving PA model p_{sum}, γ and Eqn. (4) take respectively the following forms:*

$$p_{sum} = \left(1 - \frac{2\sqrt{k}}{\sqrt{2\pi}(m-1)}\right)\theta + \frac{2\sqrt{k}}{\sqrt{2\pi}(m-1)},$$

$$\gamma = \frac{1}{\sqrt{2\pi(k-1)}}\left(\frac{k}{2m} + \frac{3}{2} + \frac{m\ln\frac{m-k}{m}}{2k} - \frac{\sqrt{t}}{2(t-1)}\right)p_{sum},$$

$$2mtY_t = 2m(2t-1)Y_{t-1} + \left(m + M\left(L - \frac{\sqrt{t}}{2(t-1)}\right)\right)((1-N)\theta + N). \tag{17}$$

If θ satisfies $\Theta\left(\frac{1}{\sqrt{n}-\ln(n-1)}\right)$, the network wide influence spreading will happen with high probability.

5.3 Cascading in Generalized k-complex contagion Model

By introducing dynamics into node's 'influence threshold', the k-complex contagion model is able to capture individual differences and thus better reflect reality. We consider in this paper two types of threshold dynamics, with the 'influence threshold' exhibiting two distinct distributions. Again, we only present the most important proof techniques involved. Details are available in the full version [31].

5.3.1 Proof of Lemma 4.4(2). We first sketch the proof when 'influence threshold' k follows uniform distribution in an evolving ER network. Thanks to the uniform distribution, we have $\sum_{R_u \sim \mathcal{D}} \Pr[R_u = k] = \sum_{k=1}^{n-1} \frac{1}{n-1}$.

Corollary 5.6. *Embedding $\sum_{R_u \sim \mathcal{D}} Pr[R_u = k]$ into Eqn. (8) and (7), we derive the following corollary by virtue of chernoff bound.*

$$Pr[I_i | N_u^i, I_u] = pp_0 \tag{18}$$

$$Pr[I_u = 1] = p_0 + (1-p_0)\left(1 - \sum_{k=1}^{n-1} \frac{1}{n-1} Pr[npp_0 + (n'-n)pp_0 \le k]\right) \tag{19}$$

In conclusion, if $n' \ge \Theta(\frac{n}{pp_0})$, when $n \longrightarrow \infty$, the scale comparable to the 'impasse' part will be influenced.

5.3.2 Proof of Theorem 4.1(2). In the evolving ER model, by assuming $n' = n$ in the above proof (i.e. there is no 'reactivation' part in the network), we need $p_0 = \Theta(1)$ proportion of influenced nodes in the network to complete the cascading during the 'inner' diffusion. Similar to previous demonstrations, we focus on determining the influence spread by 'initial' diffusion in the network.

By specifying Eqn. (5) and (6), the following corollary gives the expressions of p_{sum} and γ .

Corollary 5.7. *In an ER network where k is uniformly distributed,*

$$p_{sum} = (1 - pY_{t-1})\theta + pY_{t-1}, \qquad \gamma = pp_{sum}. \tag{20}$$

Substituting the above recursive equations into Eqn. (4), we obtain:

$$tY_t = \left(t - 1 + \frac{t-1+p}{t-1}p - \frac{t-1+p}{t-1}p\theta\right)Y_{t-1} + \frac{t-1+p}{t-1}\theta. \tag{21}$$

Corollary 5.8. *When network size is t, the proportion of the 'influenced' nodes in the network (i.e.Y_t) is*

$$Y_t = \frac{\theta}{t^{1-p(1-\theta)}} + \frac{\theta}{1-p(1-\theta)} \tag{22}$$

As $t = n \longrightarrow \infty$, $\theta = \frac{\Theta(1-p)}{1-\Theta(p)}$.

5.3.3 Proof of Lemma 4.4(3). This Lemma proves the reactivation phenomenon under ER model when k obeys the Poisson Distribution. By showing the power of nodes in the 'reactivation' part and revealing the influence spreading via 'inner' diffusion among nodes in the 'impasse' part, it find the required number of nodes in the 'reactivation' part (i.e. $n' - n$). The analysis is similar to that of Lemma 4.4 (2), and we only present here a sketch of the proof.

Integrating Poisson distribution into Eqn. (7) and (8), we have:

$$\Pr[I_i | N_u^i, I_u] = \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(n-1)} \tag{23}$$

$$\Pr[I_u = 1] = 1 - \sum_{k=1}^{n-1} e^{-\lambda} \frac{\lambda^k}{k!} \Pr\left[npp_0 + (n'-n)\frac{2\sqrt{\lambda}}{\sqrt{2\pi}(n-1)}\right] \tag{24}$$

The probability that the entire network is influenced is:

$$\Pr[\text{First } n \text{ number of nodes get influenced}] =$$
$$1 - \frac{n(1-p_0)e^{-\lambda}e^{\lambda e^{t_1}}}{\exp\left(npp_0 + (n'-n)\frac{2\sqrt{\lambda}}{\sqrt{2\pi}(n-1)}\right)t_1} \tag{25}$$

where t_1 is a constant parameter of chernoff bound. In conclusion, when $npp_0 + (n'-n)\frac{2\sqrt{\lambda}}{\sqrt{2\pi}(n-1)} \ge 2\ln n + \lambda(e^{t_1} - 1)$ is satisfied, it is almost certain that all nodes in the 'impasse' part are influenced.

5.3.4 Proof of Theorem 4.1(3). Similarly to the former proof, we let $n' = n$ (i.e. omit the 'reactivation' part), and find that we need roughly $\Theta(2\ln n)$ seeds to complete the cascading during the 'inner' diffusion. In the following we reveal the power of 'initial' diffusion under Poisson distribution in an evolving ER network and find the required θ.

Corollary 5.9. *Embedding the Poisson distribution into Eqn. (7) and (8), we have:*

$$p_{sum} = \left(1 - \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(t-1)}\right)\theta + \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(t-1)} \text{ and } \gamma \approx \frac{(t-1)p^2 p_{sum}}{\sqrt{2\pi}}$$

The recursive equation of Y_t is:

$$tY_t = (t-1)Y_{t-1} + (1 + \frac{p}{\sqrt{2\pi}})\left(\left(1 - \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(t-1)}\right)\theta + \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(t-1)}\right)$$

Corollary 5.10. *When the network size tends to $n \longrightarrow \infty$ and $p_0 = \Theta\left(\frac{\ln n}{n}\right)$, we need to choose $\theta = \Theta\left(\frac{\ln\left(\frac{\sqrt{\lambda}n}{n-1}\right)}{n}\right)$ proportion of nodes in the network to be the seeds.*

5.3.5 Proof of Theorem 4.5(2) and (3). The reasoning differences under two generalized k-complex contagion models in an evolving PA network mainly hide in the mathematical process. Here we just present the key results during the proof by combining Theorem 4.5 (2) and (3) together with the help of Lemmas 5.11 and 5.12.

Lemma 5.11. *Suppose $n \longrightarrow \infty$. When 'influence threshold' k follows the Uniform Distribution in an evolving PA network, $p_0 = \Theta\left(\frac{\sqrt{n}}{n}\right)$. Let $H = \frac{(\sqrt{m-1}-1)}{m^2\sqrt{2\pi}}$ $(n >> H)$. For each newly coming node:*

$$p_{sum} = (1 - Y_{t-1})\theta + Y_{t-1} \text{ and } \gamma = \frac{2(\sqrt{m-1}-1)}{m\sqrt{2\pi}} p_{sum}$$

Table 1: Theoretical Fitting Curve

Situation	Probability	Fitting Result
$G \in \text{ER} \ \& \ \mathscr{D} \in \text{Fixed}$	$\theta = a \cdot \frac{2\ln n - D(\ ln(\frac{\sqrt{2\pi}}{2\sqrt{k}}(n-1)) - E)}{An - B\ln(n-1) + C}$	$a = 4.903$
$G \in \text{ER} \ \& \ \mathscr{D} \in \text{Uniform}$	$\theta = a \cdot \frac{1-p}{1-bp}$	$a = 0.96, b = 2.10$
$G \in \text{ER} \ \& \ \mathscr{D} \in \text{Poisson}$	$\theta = a \cdot \frac{4\ln n - (1+\frac{1}{\sqrt{2\pi}})\frac{\sqrt{2\pi}}{2\sqrt{\lambda}}\ln(\frac{\sqrt{2\pi}}{2\sqrt{\lambda}}(n-1))}{n - \frac{2\sqrt{\lambda}}{\sqrt{2\pi}}\ln(n-1) + 1}$	$a = 13.16$
$G \in \text{PA} \ \& \ \mathscr{D} \in \text{Fixed}$	$\theta = a \cdot \frac{1}{\sqrt{t} - ln(t-1)}$	$a = 0.43$
$G \in \text{PA} \ \& \ \mathscr{D} \in \text{Uniform}$	$\theta = a \cdot \frac{1}{\sqrt{t}(t+H)^H}$	$a = 0.014$
$G \in \text{PA} \ \& \ \mathscr{D} \in \text{Poisson}$	$\theta = a \cdot \frac{1}{\sqrt{t} - ln(t-1)}$	$a = 0.6031$

[1] A, B, C, D, E are all constants. The specific details of them could be seen in Corollary 5.3.

Then we can find that the seed proportion required for a network-wide influence is $\theta = \Theta\left(\frac{1}{\sqrt{n}(n+H)^H}\right)$

Lemma 5.12. *When 'influence threshold' k follows the Poisson Distribution in an evolving PA network,* $p_0 = \Theta\left(\frac{\sqrt{n}}{n}\right)$. *For each newly coming node, we have:*

$$p_{sum} = \left(1 - \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(m-1)}\right)\theta + \frac{2\sqrt{\lambda}}{\sqrt{2\pi}(m-1)}$$

$$\gamma = \frac{1}{\sqrt{2\pi}}\left(\frac{\sqrt{\lambda}}{2m} + \frac{3}{2\sqrt{\lambda-1}} + M - \frac{\sqrt{t}}{2\sqrt{\lambda-1}(t-1)}\right)p_{sum}$$

where M is a constant with expression $\sum_{k=1}^{m} e^{-\lambda}\frac{\lambda^k}{k!}\frac{m\ln\frac{m-k}{m}}{2k}$. *Then we can get the result that the proportion required for a network-wide influence is* $\theta = \Theta\left(\frac{1}{\sqrt{n} - \ln(n-1)}\right)$.

6 EXPERIMENTS

We perform simulations under both synthetic and real network to perform the principle of information cascading in evolving networks and to verify our estimation of the seed quantity order under each setting. The accuracy of theoretic estimations is illustrated by position of simulated curves and actual curves.

6.1 Experiments under Synthetic Networks

For synthetic networks, we generate graphs with size ranging from 10000 to 500000 by using separately the evolving ER model $G(n, p)$ and the evolving PA model. We adopt $p = 0.03$ for all evolving ER construction and we set $m = 10$ for evolving PA model $PA_{m,n}(V, E)$. And we take $k = 5$ in traditional k-complex model, while $\lambda = 10$ in Poisson Distribution. In order to get closer to reality and to increase the extensiveness of our work, we consider in addition the Gaussian Distribution as the third possible distribution of k. Analysis under the Gaussian model is too difficult, therefore we only studied the case empirically and observe the information cascading during the experiments. We set $\mu = 8, \delta = 2$ and discretized the Gaussian Distribution by $p(x) = F(x + 0.5) - F(x - 0.5)$ and normalized the sum of the probability to 1.

Now we would like to present one by one the information cascading model setting and the fitting equations in ER networks and PA networks. In each fitting equation, θ is the actual proportion of seeds in the whole network and a is the fitting parameter of the theoretic formula. The results are presented in Table [1]. The parameter 'a'(and 'b') in each situation is generated by curve-fitting the theoretical estimation with the first four nodes.

Figure 2 (a) shows the results where the 'influence threshold' is a constant. Our experiment shows that once determined the parameter 'a' by first four nodes, theoretic estimations match perfectly the actual seed numbers under both network topologies, which shows that our result is a critical bound of seeds once we choose them randomly. Alos, we find that less seeds are needed in larger ER networks. This seemingly counter-intuitive result can be justified by the fact that late birds emit much more edges than early ones in ER networks and greater connectivity is intuitively prone to information cascading.

Figure 2 (b) manifests the accuracy of our theoretic estimations in ER networks where the 'influence threshold' k has certain dynamics. The overlapping horizontal lines are zoomed for a better discrimination. Apart from the case where k follows the uniform distribution, the seed quantity decreases as the network expands. The reason that the required seed number rises accordingly with network size under the uniformly distributed k is that there are quite a few 'unbending' nodes with high 'influence thresholds', hampering the information cascading.

Figure 2 (c) illustrates the correctness of our theoretic estimations in PA networks where the 'influence threshold' k has certain dynamics. The overlapping horizontal lines are zoomed for a better discrimination. The results in general is quite the opposite of those obtained in ER networks. Much less seeds are needed when the 'influence threshold' follows the uniform distribution than the poisson distribution or normal distribution. As λ and μ are close to m, there are more nodes with high 'influence threshold's and hence a higher average of k under the poisson or normal distribution. The results indicates the impact of R_u's value on the necessary seed number.

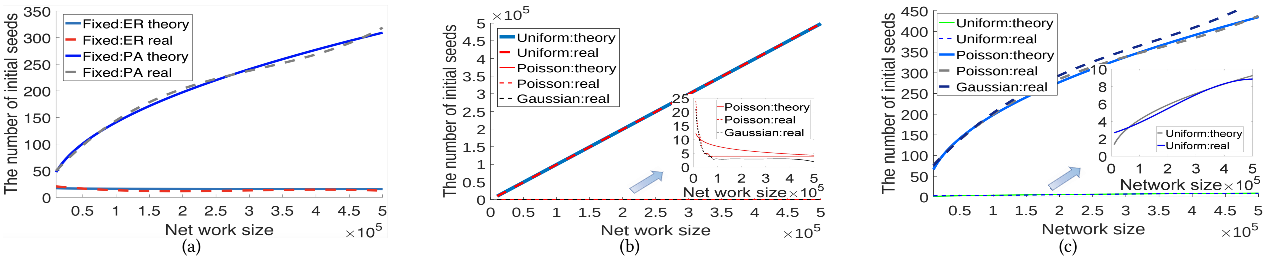
6.2 Experiments under Real Networks

Our real networks are the coauthor datasets of machine learning and bioinformatics from 1965 to 2016. As it has been observed in [28] that the both networks comply with the PA model, the experiments in this part can be seen as the experiments under the PA network topology under uniform 'influence threshold', with $k = 5$.

Since the annual growth of networks is small, we decide instead to classify the data by longer time intervals. We divide the datasets by six time stamps: 1980, 1990, 2000, 2005, 2010 and 2016. Due to the accelerating growth speed of these networks after 2000, we cut the time interval by half(from 10 years to 5 years). The machine learning network has 1.51 million nodes and the bioinformatics network contains altogether 1.82 million nodes.

In real networks, there is one new node coming at every time slot. The nodes that arrive in the same year are added with a random sequential order. In cases where there comes a new node which does not have any connections to the existing nodes or has less than 10 edges connect with former nodes, we will manually add the number of their edges up to 10 according to the PA model so that the newcomer is not isolated from other nodes.

Figure 3 shows the results under the machine learning network and the bioinformatics network. The simulated curve is drawn with the theoretic estimation $n\theta = a \cdot n \cdot \frac{1}{\sqrt{t} - ln(t-1)}$ as both networks are characterised by a power-law degree distribution. The real data points are scattered evenly above and below the simulated curve,



(a)

(b)

(c)

Figure 2: Relation between network size and seed number under (a) traditional k-complex model with k=5, (b) simulated ER graphs and (c) simulated PA graphs.

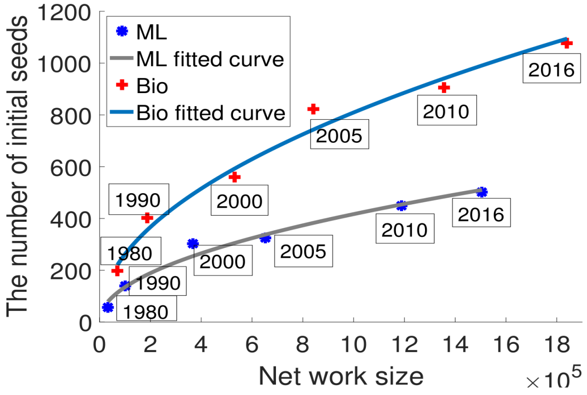


Figure 3: Relation between network size and seed number in real networks

confirming the PA network topology of the two real networks. Once again, our theoretic evaluation has excellent performances, with fitting parameter a being 0.4111 in machine learning network and 0.7982 in bioinformatics network.

7 CONCLUSION

This paper initiates the study of cascading under different evolving models with generalized influence thresholds. The theoretical estimations and the experiments both illustrate the counter-intuitive fact that the proportion of the required number of initial seeds decreases as the network evolves in the ER model, which firmly demonstrates the information diffusion capability of late birds. And the result in PA model shows we could use some later nodes with low degree to take the place of early birds. The extension of the problem could aim at discussing the results on a more general network combining the ER and PA model together, or consider the situation when influence diffusion in the network is not infinity.

REFERENCES

[1] R. Mermelstein, S.Cohen, E. Lichtenstein, J. S. Lichtenstein, and T. Kamarck. Social Support and Smoking Cessation and Maintenance. Journal of Consulting and Clinical Psychology, 54(4), 447-453, 1986.

[2] A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson. The Diffusion of Microfinance. Science 341, 1236498 (2013).DOI: 10.1126/science.1236498

[3] L. A. Adamic, and N. Glance. The Political Blogosphere and The 2004 US Election: Divided They Blog. In Proceedings of the 3rd International Workshop on Link discovery. 36-43, 2005.

[4] https://techcrunch.com/2017/06/27/facebook-2-billion-users/

[5] E. Breza, A. G. Chandrasekhar, T. H. McCormick and M. Pan, “Using aggregated relational data to feasibly identify network structure without network data”, in arXiv preprint arXiv:1703.04157, 2017.

[6] M. Akbarpour, S. Malladi and A. Saberi, “Diffusion, Seeding, and the Value of Network Information”, in ACM conference on Economics and Computation (EC) 2018. Revise and Resubmit, American Economic Review (AER).

[7] H. T. Nguyen, T.i P. Nguyen, T. N. Vu, and T. N. Dinh. Outward infuence and cascade size estimation in billion-scale networks. In Proc. Sigmetrics. ACM, 2017.

[8] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck. Sketch-based infuence maximization and computation: Scaling up with guarantees. In Proc. CIKM, pages 629fi??638. ACM, 2014.

[9] J. S. Coleman, E. Katz, and H. Menzel. Medical Innovation: A Diffusion Study. Bobbs-Merrill Co., 1966.

[10] J. S. Macdonald, and L. D. Macdonald. Chain Migration, Ethnic Neighborhood Formation and Social Networks. The Milbank Memorial Fund Quartly 42,1(1964),82-97, 1964.

[11] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, fiStruc- tural diversity in social contagion,fi Proc. National Academy of Sciences, vol. 109, no. 16, pp. 5962fi??5966, 2012.

[12] D. M. Romero, B. Meeder, and J. Kleinberg, fiDifferences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter,fi in Proc. of the 20th conference on World Wide Web, 2011, pp. 695fi??704.

[13] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 44-54, 2006.

[14] K. Han, C. Xu, F. Gui, S. Tang, H. Huang, and J. Luo. Discount Allocation for Revenue Maximization in Online Social Networks. ACM MobiHoc 2018 https://dl.acm.org/citation.cfm?id=3209595

[15] J. Gao, G. Ghasemiesfeh, G. Schoenebeck, and F. Yu. General Threshold Model for Social Cascades: Analysis and Simulations. In Proceedings of the 2016 ACM Conference on Economics and Computation (EC ’16). ACM, New York, NY, USA, 617-634. DOI:https://doi.org/10.1145/2940716.2940778

[16] P. Erds, and A. Rnyi. On the Evolution of Random Graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences. vol.5, 1960.

[17] A. L. Barabsi, and R. Albert. Emergence of Scaling in Random Networks. Science,286(5439):509-512, 1999.

[18] N. Berger, C. Borgs, J. T. Chayes, and A. Saberi, On The Spread of Viruses on The Internet. In *Proc. of The 16th ACM-SIAM Symposium on Discrete Algorithms*, 301-310, 2005.

[19] B. A. Prakash, D. Chakrabarti, N. Valler, M. Faloutsos, and C. Faloutsos. Threshold Conditions for Arbitrary Cascade Models on Arbitrary Networks. IEEE IDCM 2011 https://faculty.mccombs.utexas.edu/deepayan.chakrabarti/mywww/papers/icdm11-threshold.pdf

[20] J. Chalupa, P. L. Leath, and G. R. Reich. Bootstrap Percolation on A Bethe Lattice. *Journal of Physics C: Solid State Physics*, vol.12, no. 1, p. L31, 1979.

[21] M. Garetto, E. Leonardi, and G. L. Torrisi. Generalized threshold-based epidemics in random graphs: the power of extreme values. ACM Sigmetrics 2016 https://arxiv.org/abs/1603.04643

[22] A. Coja-Oghlan, U. Feige, M. Krivelevich, and D. Reichman. Contagious Sets in Expanders. *Proc. of The 26th ACM-SIAM Symposium on Discrete Algorithms*, 1953-1987, 2015.

[23] R. Ebrahimi, J. Gao, G. Ghasemiesfeh, and G. Schoenbeck. How Complex Contagions Spread Quickly in Preferential Attachment Models and Other Time-Evolving Networks. arXiv preprint arXiv:1404.2668, 2014.

[24] G.Tong, W.Wu, S.Tang, and D.Du. Adaptive Infuence Maximization in Dynamic Social Networks. *IEEE/ACM Transactions on Networking (TON)*, 2017.

[25] Z. Wen, B. Kveton, and M. Valko. Infuence Maximization with Semi-Bandit Feedback. arXiv preprint arXiv:1605.06593, 2016.

[26] S. Lei, S. Maniu, L. Mo, R. Cheng, and P. Senellart. Online Influence Maximization. In *Proc. SIGKDD*, pages 645fi??654. ACM, 2015.

[27] Albert, R & Barabasi, Albert-Laszlo. (1999). Emergence of Scaling in Random Networks. 286. 1-11.

[28] X. Wu, L. Fu. , J. Meng. and X. Wang. Evolving Influence Maximization. https://arxiv.org/abs/1804.00802.

[29] Erdös, P.; Rényi, A. (1959). ”On Random Graphs. I” Publicationes Mathematicae.6: 290-297

[30] Bollobás, B. (2001). Random Graphs (2nd ed.). Cambridge University Press.

[31] https://www.dropbox.com/s/emvqo8jpih21wfa/Complete_MobiHoc.pdf?dl=0