On Maximizing Diffusion Speed in Social Networks: Impact of Random Seeding and Clustering *

Jungseul Ok, Youngmi Jin, Jinwoo Shin, and Yung Yi
Department of Electrical Engineering, KAIST
Daejeon, Republic of Korea
{ockjs, youngmi_jin, jinwoos, yiyung}@kaist.ac.kr

ABSTRACT

A variety of models have been proposed and analyzed to understand how a new innovation (e.g., a technology, a product, or even a behavior) diffuses over a social network, broadly classified into either of epidemic-based or game-based ones. In this paper, we consider a game-based model, where each individual makes a selfish, rational choice in terms of its payoff in adopting the new innovation, but with some noise. We study how diffusion effect can be maximized by seeding a subset of individuals (within a given budget), i.e., convincing them to pre-adopt a new innovation. In particular, we aim at finding 'good' seeds for minimizing the time to infect all others, i.e., *diffusion speed maximization*. To this end, we design polynomial-time approximation algorithms for three representative classes, Erdős-Rényi, planted partition and geometrically structured graph models, which correspond to globally well-connected, locally well-connected with large clusters and locally well-connected with small clusters, respectively, provide their performance guarantee in terms of approximation and complexity. First, for the dense Erdős-Rényi and planted partition graphs, we show that an arbitrary seeding and a simple seeding proportional to the size of clusters are almost optimal with high probability. Second, for geometrically structured sparse graphs, including planar and d-dimensional graphs, our algorithm that (a) constructs clusters, (b) seeds the border individuals among clusters, and (c) greedily seeds inside each cluster always outputs an almost optimal solution. We validate our theoretical findings with extensive simulations under a real social graph. We believe that our results provide new practical insights on how to seed over a social network depending on its connection structure, where individuals rationally adopt a new innovation. To our best knowledge, we are the first to study such diffusion speed maximization on the game-based diffusion, while the extensive research efforts have been made in epidemic-based models, often referred to as *influence maximization*.

*This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2013R1A2A2A01067633, NRF-2013R1A1A3A04007104)

SIGMETRICS'14, June 16–20, 2014, Austin, Texas, USA.

http://dx.doi.org/10.1145/2591971.2591991.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Algorithms, Theory, Performance

Keywords

Influence Maximization, Clustering, Random Seeding

1. INTRODUCTION

People are actively using social networks to get new information, exchange new ideas or behaviors, and adopt new innovations. Clearly, it is of significant importance to understand how such information diffuses over time, where diffusion by local interaction is the most prominent feature. Various fields including computer science, economics, and sociology have expressed their interests in understanding diffusion, e.g., [10, 40, 42]. People have first started to propose diffusion models in social network with close relevance to studies with long history on raging epidemic, e.g., SIRS model [26] or interacting particle system, e.g., Ising model [20]. Examples of such epidemic-based diffusion model also include [15] and [6], often referred to as independent cascade or linear threshold models [24].

Different from epidemic-based models, people often make strategic choices, i.e., an individual adopts a new technology only if the new technology provides sufficient utility, which increases with the number of neighbors who adopt the same technology (i.e., coordination effect) [14, 19, 33, 35]. This is called game-based diffusion model, which is the main focus of this paper. A recent work by Montanari and Saberi [33] addressed the question of the equilibrium behavior as well as the impact of topological properties on convergence speed. Under the assumption that individuals behave with bounded rationality (i.e., noisy best response dynamic), it has been proved that the number of innovation adopters increases and the innovation finally becomes widespread. However, the diffusion time can be significantly long so that in practice the innovation often diffuses within only a small number of individuals or even become extinct in practice. One of the approaches to reduce the convergence time is to *seed* some individuals, i.e., convince a subset of individuals to pre-adopt the new innovation, e.g., by providing some incentives to those users.

The problem of maximizing the "degree of diffusion" by properly selecting seeds has been popularly studied in epidemic-based models, often referred to as *influence maximization*, whose major goal is to maximize the number of infected individuals. However,

in game-based models, as in e.g., [33], the problem becomes completely different mainly because diffusion is widespread at the equilibrium. Thus, we study how to choose a constrained set of individuals to accelerate the speed of diffusion, which we call *diffusion speed maximization*.

1.1 Contribution

We first formulate a diffusion speed maximization problem, say ***P1***, as minimizing the notion of typical hitting time which measures the time when every individual adopts the innovation. We discuss its computational challenges mainly stemming from (i) MCMC (Markov Chain Monte Carlo) based estimation and (ii) probabilistic feature of a typical hitting time, which is neither algebraic nor combinatorial (see Section 2.3). Therefore, we transform the original problem ***P1*** into a combinatorial optimization, say ***P2***, using the theory of meta-stability of Markov chains [39], which, however, turns out to be computationally intractable as well as difficult to be reduced to a classical NP-hard problem amenable to approximation. For example, the influence maximization in epidemic-based models becomes the submodular maximization in most cases, whose greedy algorithm guarantees constant approximation [24]. However, we found that the optimization ***P2*** is not a submodular problem (see our discussion in Section 2.3).

Despite this hardness of ***P2***, we propose polynomial-time approximation algorithms for three graph classes, Erdős-Rényi, planted partition and geometrically structured graphs, and their provable performance guarantees in terms of approximation ratio as well as complexity. Our contribution lies in providing new insights on how to seed individuals depending on the connection structure of underlying graph topologies.

- ***Erdős-Rényi and planted partition graphs.*** We show that an arbitrary seeding and a simple seeding proportional to the size of clusters are close to an optimal one with high probability for the dense Erdős-Rényi and planted partition graphs, respectively (see Theorems 3.1 and 3.2). The main technical ingredient for this result is on our concentration inequalities on the so-called 'energy function' (see Lemma 4.1), which provides the exact approximation qualities of the random seeding via a solution of certain quartic equations. Then it is provably almost optimal via obtaining its approximate close-form solution.
- ***Geometrically structured graphs.*** For this graph class, including planar and d-dimensional graphs, we design an algorithmic framework, called PaS (Partitioning and Seeding), and provide a condition, which, if met, provably guarantees good approximation with polynomial complexity (see Theorem 3.3). PaS consists of two phases: (i) partitioning the graph into multiple clusters, and (ii) seeding within each cluster. The proposed PaS framework relies on our finding that the diffusion process in a graph is dominated by the slowest diffusion process among the underlying clusters. Thus, in the partitioning phase, a given graph should be smartly partitioned into the clusters in which a seeding problem becomes tractable (via seeding the "border individuals" among clusters). Then, to minimize the diffusion time, our focus simply becomes a good seed budget allocation to each cluster that minimizes the overall diffusion time. A greedy algorithm is run to achieve the desired budget allocation in the seeding phase.

The practical implications from our theoretical findings are summarized in what follows: Erdős-Rényi, planted partition and geometrically structured graphs represent *(a)* globally well-connected, *(b)* locally well-connected with big clusters, and *(c)* locally well-connected with small clusters, respectively. First, for globally well-connected graphs like Erdős-Rényi graphs, careful seeding is not highly required, because the underlying topological structure such as high symmetry and connectivity does not change significantly even after seeding with a small budget. However, for locally well-connected graphs, it is necessary to intelligently exploit their clustering characteristics, where the network-wide diffusion time is governed by both intra-cluster diffusion and inter-cluster correlation. As is in sharp contrast to epidemic-based models, in game-based ones, it turns out that in *(b)* intra-cluster diffusion becomes the dominant factor, as opposed to in *(c)* where inter-cluster correlation dominantly determines the network-wide diffusion speed. Thus, as described in Sections 3.2 and 3.3, for planted partition graphs, we focus only on how to distribute the seed budget to each (big) cluster, while for geometrically structured graphs, the seeds are mainly selected from the border individuals to remove inter-cluster correlation.

1.2 Related Work

As discussed earlier, diffusion models in literature can be broadly classified into: (i) epidemic-based [2–4,13,17,24,26] and (ii) game-based [5, 14, 23, 43], depending on how diffusion occurs, i.e., just like a contagious disease or individuals' strategic choices. In particular, game-based diffusion models [5, 14, 23, 43] adopt a networked coordination game where the payoff matrix appropriately models the value of accepting new technology for the neighbors' selections, and studied the equilibrium and the dynamics. Especially, Kandori et al. [23] proved that the noisy best response dynamic converges to the equilibrium that the innovation becomes widespread. Recently, significant attention has been paid to the study of convergence time. In [33], it was shown that in highly connected graph, the convergence becomes slower as opposed to in epidemic models. In [21], the authors showed that the external information such as advertisement on a new technology may slow down diffusion, again on the contrary to in epidemic models [4]. In practice, a small set of influential nodes, called *seeds*, can be convinced to pre-adopt a new technology, which can increase the effect of diffusion. See [12] for motivation in viral marketing, [37] in graph detection, and [27] in computer virus vaccine dissemination. The problem of how to maximize the diffusion effect for both diffusion models are summarized next, where depending on the adopted diffusion model, different problems can be formulated.

Epidemic-based model. In [24, 25], the authors addressed the so-called influence maximization problem in linear threshold (LT) and independent cascade (IC) models. In both LT and IC models, each individual has only one chance to infect its neighbors right after its infection. Thus, a main goal is to maximize the influence spread, i.e., maximize the number of infected individuals. In [24,25], it was first discussed that the problem is computationally intractable because of #P-completeness in measuring influence spread for a given seed set and NP-completeness in finding the optimal seed set that maximizes influence spread. Using the technique on the submodular set function maximization in [36], they showed that a greedy algorithm achieves at least $(1 - 1/e - \varepsilon)$ of the optimal influence spread where ε represents the inaccuracy of Monte Carlo simulation for measuring the influence spread. Since the Monte-Carlo based measurement does not tend to scale with the network size, the authors in [9] proposed a scalable method called MIA using a tree structure. In [18], a clustering concept is proposed to reduce the computational complexity in measuring the influence spread. In [8], Chen et al. proposed modified LT and IC models by adding contact process, which delays infection chance of the infected individual from its infection. Using the modified models, the authors formulated an influence maximization with time deadline and pro-

posed a greedy algorithm motivated by [24, 25]. In [16], Goyal et al. generalized the influence maximization problem in LT and IC models as an optimization problem with three dimensions: influence spread, seed budget, and time deadline.

Game-based model. In [11, 23, 28, 35], the authors considered only the best-response dynamics and studied the conditions (of network topology and the payoff difference between old and new technologies) on the existence of a small seed set, referred as the so-called "contagion set," under which all individuals adopt new technology. In [29], a noisy best response was considered with objective of maximizing the influence spread by choosing a seed set assuming that there exists a set of "negative individuals," and a greedy algorithm was proposed with simulation-based evaluations. As discussed in [33], without negative seeding, it is guaranteed to converge to a state where all individuals adopt the new technology. This paper studies a problem of minimizing the convergence time to such an equilibrium under a noisy best response dynamic. To the best of our knowledge, this paper is the first to study this diffusion speed maximization in a game-based diffusion model.

2. MODEL AND FORMULATION

2.1 Network Model and Coordination Game

Network model. We consider a social network as an undirected graph $G = (V, E)$, where V is the set of n nodes and E is the set of edges. Each node represents an individual (or a user) and each edge represents a social relationship between two individuals. We let $N(i)$ be the set of node i's neighbors, i.e., $N(i) = \{j \in V \mid (i, j) \in E\}$. We simply use +1 and -1 to refer to new and old technologies, respectively. We are interested in how a new technology diffuses over the network.

Networked coordination game. We first consider the famous two-person coordination game whose payoff matrix is given by Table 1, where an individual can choose one of new or old technologies, +1 and -1. We make the following practical assumptions on the payoffs. First, there always exists coordination gain, i.e., $a > d$ and $b > c$. Second, coordination gain becomes larger for the new technology, i.e., $a - d > b - c$.

Table 1: Two-person coordination game

P	$+1$	-1
$+1$	(a, a)	(c, d)
-1	(d, c)	(b, b)

The two-person coordination game is extended to an n-person game over G. We let $\boldsymbol{x} = (x_j \in \{-1, +1\} : j \in V)$, and $\boldsymbol{x}_{-i} = (x_j : j \in V \setminus \{i\})$ be the states (i.e., a strategy vector chosen by the entire nodes) of all and those except for i, respectively. Then, in n-person game over G, node i's payoff $P_i(x_i, \boldsymbol{x}_{-i})$ for the state $\boldsymbol{x}$ is modeled to be the aggregate payoff against all of i's neighbors, i.e.,

$$P_i(x_i, \boldsymbol{x}_{-i}) \quad = \quad \sum_{j \in N(i)} P(x_i, x_j), \tag{1}$$

where $P(x_i, x_j)$ is the payoff from the two-person coordination game, as in Table 1. For notational convenience, let $-\mathbf{1} =$ (resp. $+\mathbf{1}$) denote the state where every user adopts -1 (resp. $+1$).

2.2 Diffusion Dynamics

Seed set. We consider a continuous time model, where each node updates its strategy whenever its own independent Poisson clock with unit rate ticks. Let $\boldsymbol{x}(t) = (x_i(t) : i \in V) \in \{+1, -1\}^V$ be the network state at time t, representing the strategies of all nodes at time t. We introduce the notion of *seed set* $C \subset V$, where each node in C is initialized by $+1$ and does not change its strategy over all time, i.e., for any $i \in C$, $x_i(t) = +1$ for all $t \geq 0$. Next, we describe how each non-seed individual updates its strategy.

Best response. As is well-known in game theory, in the best response dynamics, each (non-seed) individual selects a strategy that maximizes its own payoff: a node i chooses $+1$, if

$$(a - d)|N^+(i)| \quad \geq \quad (b - c)|N^-(i)| \tag{2}$$

where $N^+(i)$ and $N^-(i)$ denote the sets of node i's neighbors adopting $+1$ and -1, respectively. Noting that for a given state $\boldsymbol{x}$ $P_i(+1, \boldsymbol{x}_{-i}) - P_i(-1, \boldsymbol{x}_{-i})$ represents the payoff difference between when node i chooses +1 and -1, the best response of node i is sign$(P_i(+1, \boldsymbol{x}_{-i}) - P_i(-1, \boldsymbol{x}_{-i}))$, simply expressed as:

$$\text{sign}\left(h_i + \sum_{j \in N(i)} x_j\right), \tag{3}$$

where $h_i = h|N(i)|$ and $h = \frac{a-d-b+c}{a-d+b-c}$

Noisy best response: Logit dynamics. In practice, individuals do not always make the "best" decision. We model such behavior by introducing small mutation probability that non-optimal strategy is chosen, often called noisy best response. A version of the noisy best response we focus on in this paper is *logit dynamics* [5, 31, 32, 34] that individuals adopt a strategy according to a distribution of the logit form which allocates larger probability to those strategies delivering larger payoffs. More formally, for the given state $\boldsymbol{x}$, non-seeded node i chooses the strategy $y_i \in \{-1, +1\}$ with the following probability:

$$\mathbb{P}_\beta(y_i|\boldsymbol{x}) = \frac{\exp(\beta y_i K_i(\boldsymbol{x}))}{\exp(\beta K_i(\boldsymbol{x})) + \exp(-\beta K_i(\boldsymbol{x}))}. \tag{4}$$

where

$$K_i(\boldsymbol{x}) = \frac{1}{2}\left(h_i + \sum_{j \in N(i)} x_j\right).$$

Note that $(a - d + b - c)y_i K_i(\boldsymbol{x})$ is the payoff gain for the strategy y_i instead of $-y_i$ from (3) and $(a - d + b - c)$ is removed just for convenient handling of other quantities later. Here, the parameter β represents the degree of user rationality, where $\beta = \infty$ corresponds to the best response and $\beta = 0$ lets users update their strategies uniformly at random. When the state changes according to the probability (4) and nodes' independent Poisson clock ticks, the system can be viewed as a continuous Markov chain with the state space $\mathcal{S}_C = \{\boldsymbol{z} \in \{-1, +1\}^V \mid z_i = 1 \text{ if } i \in C\}$, recall C is a given seed set. The dynamics here is also called the Glauber dynamics in the "truncated" Ising model [38], where the truncation occurs due to the existence of hard-coded nodes (i.e., the nodes in the seed set C). Then, it is not hard to see that this chain is time-reversible with the following stationary distribution μ_β:

$$\mu_\beta(\boldsymbol{x}) \propto \exp(-\beta H(\boldsymbol{x})),$$

where

$$H(\boldsymbol{x}) = -\frac{1}{2}\left\{\sum_{(i,j) \in E} x_i x_j + \sum_{i \in V} h_i x_i\right\} + (1 + 2h)|E|. \tag{5}$$

In the above, the constant term $(1 + 2h)|E|$ is not necessarily needed to characterize the stationary distribution, but we add due to notational convenience in our proofs. We note that $-H$ is often referred to as a *potential* function of the n-person game described in Section 2.1 and H is called the *energy* function in literature.

2.3 Problem Formulation

Our objective is to find a seed set C (within some budget constraint) which maximizes the speed of diffusion. To this end, we define a couple of related concepts.

First, a random variable called the hitting time (to the state where all users adopt $+1$) of our system with a seed set C starting from the initial state $\boldsymbol{y} \in \mathcal{S}_C$ defined by:

$$T_+(C, \boldsymbol{y}) = \inf\{t \geq 0 \mid \boldsymbol{x}(t) = +\mathbf{1},\ \boldsymbol{x}(0) = \boldsymbol{y}\}.$$

Using this, we next define the *typical hitting time* to be:

$$\tau_+(C) = \sup_{\boldsymbol{y} \in \mathcal{S}_C} \inf\left\{t \geq 0 \mid \mathbb{P}_\beta\{T_+(C, \boldsymbol{y}) \geq t\} \leq e^{-1}\right\}.$$

This means that with probability $1 - 1/e$ ($> 1/2$), every node adopts the innovation $+1$ within time $\tau_+(C)$. This typical hitting time has also been used to measure the diffusion speed for a similar model via close relation between hitting and mixing of the Markov chain, e.g., see [33]. Our goal is to solve the following optimization problem:

$$\textbf{\textit{P1.}} \qquad \min_{C \subset V} \quad \tau_+(C) \qquad \text{subject to } |C| \leq k,$$

where k is the given seed budget.

Computational challenges of P1. First, given a seed set C, the computation of the typical hitting time $\tau_+(C)$ is a highly non-trivial task, primarily because the hitting time $T_+(C, \cdot)$ is a random variable decided by the Markov chain of the logit dynamics whose underlying space is exponentially large, i.e., $|\mathcal{S}_C|$. One can use the Markov Chain Monte Carlo (MCMC) method for estimating $\tau_+(C)$, which, however, takes at least the mixing time of the Markov chain of the logit dynamic that is typically exponentially large [33]. Even worse, a naive exhaustive search for the optimization ***P1*** requires computing the typical hitting time $2^{\Omega(n)}$ times for $k = \Omega(n)$. Second, the hardness of the optimization ***P1*** also comes from the probabilistic definition of the minimizing objective $\tau_+(C)$, which is neither algebraic nor combinatorial. Due to these reasons, at a first glance, the optimization ***P1*** is a highly challenging computational task, similarly to other influence maximization problems in epidemic-based diffusion models, e.g., see [24]. It is not even clear whether the decision version of the optimization ***P1*** is in the computational class NP.

Problem formulation via a combinatorial optimization. To overcome such difficulties, we use the known combinatorial characterization of the typical hitting time $\tau_+(C)$ from the theory of metastability [33, 39], where it was proved that for a given seed set $C \subset V$,

$$\tau_+(C) = \exp(\beta\Gamma^*(C) + o(\beta)), \qquad \text{as} \quad \beta \to \infty., \qquad (6)$$

where we refer to $\Gamma^*(C)$ as the *diffusion exponent* with respect to the seed set C. In the above, $\Gamma^*(C)$ is defined as

$$\Gamma^*(C) = \max_{w_0 \in \mathcal{S}_C} \min_{\underline{w}: w_0 \to +\mathbf{1}} \max_{t < |\underline{w}|} [H(w_t) - H(w_0)]. \qquad (7)$$

where the minimization is taken over every possible path $\underline{w} = (w_0, w_1, \cdots, w_T = +\mathbf{1})$ such that for each t, w_t and w_{t+1} are same except for one coordinate. This implies that Γ^* dominates the exponent of diffusion time $\tau_+(C)$ for large β. Also, Γ^* can be interpreted as the "energy barrier" along the most probable path to $+\mathbf{1}$. Two maximums in (7) choose the largest energy difference along a path toward $+\mathbf{1}$. Then the (middle) minimum in (7) finds a path that has the smallest energy barrier to the ground state $+\mathbf{1}$ so that it is the most probable. In [33], it is known that the minimization of (7) is achieved just at a *monotone* path $w_0 \prec w_2 \cdots \prec w_T$, i.e., a user is not allowed to take back from $+1$ to -1.

The formula (6) provides a tractable approach for bounding $\tau_+(C)$ through $\Gamma^*(C)$ and motivated by this, we will focus on the following optimization instead of ***P1***:

$$\textbf{\textit{P2.}} \qquad \min_{C \subset V} \quad \Gamma^*(C) \qquad \text{subject to } |C| \leq k,$$

where it becomes identical to ***P1*** as $\beta \to \infty$ from (6).

Further challenges of P2. Note that it is still challenging to compute $\Gamma^*(C)$ for a given seed set C for the following two reasons.

- First, there exist exponentially many monotone paths to consider for the minimization in (7). Characterizations of $\Gamma^*(C)$ using 'tilted cut' and 'tilted cut-width' are known, but they are also computationally intractable, e.g., see Section 4.2 of [33]. Nevertheless, $\Gamma^*(C)$ is defined as a form of combinatorial optimization and potentially more amenable to theoretical analysis than $\tau_+(C)$.
- Second, in epidemic-based diffusion models, the influence maximization problem [24], which maximizes the number of infected individuals, could enjoy an algorithmic convenience because of the key feature the objective function turns out to be submodular. Similar convenient features may also be applied to our case, which, if so, would facilitate our analysis significantly. However, unfortunately our objective function $\Gamma^\star(\cdot)$ is neither supermodular nor submodular, as proved by a counter-example in Appendix A, which motivates our study of a different kind of approximation techniques.

3. MAIN RESULT

In this section, we describe our polynomial-time approximation algorithms for the seeding problem ***P2***. Each algorithm provides the guideline on which nodes should be seeded for fast diffusion over a game-based diffusion model for each of three graph classes, which is classified by the criterion on how globally and locally well-connected nodes are. To this end, we first introduce the following notion of "approximate solution".

DEFINITION 3.1. *A seed set $C \subset V$ with $|C| \leq k$ is called a (γ, δ)-approximate solution of the seeding problem **P2** if*

$$\Gamma^*(C) \ \leq \ \gamma \cdot \min_{C' : |C'| \leq \delta k} \Gamma^*(C'),$$

where $\gamma \geq 1$ and $\delta \leq 1$.

The parameters γ and δ measure the quality of an approximation solution, quantifying the degrees of suboptimality in *objective value* and *budget*, respectively. One can observe that the solution with $(\gamma, \delta) = (1, 1)$ corresponds to an optimal solution. In what follows, we present the characteristics of approximate solutions in three graph classes which have different topological structures in terms of connectivity and the degree of clustering.

3.1 Erdős-Rényi Graphs

We first consider the popular *Erdős-Rényi (ER) graph,* denoted by $G_{\text{ER}}(n, p)$, which is a random graph on n nodes such that every node pair has an edge with probability p. Let $\lambda = np$, roughly corresponding to the average number of neighbors per node, where our focus is when $\lambda = \Omega(1)$. For ER graphs, we obtain the following result, whose proof is presented in Section 4.1.

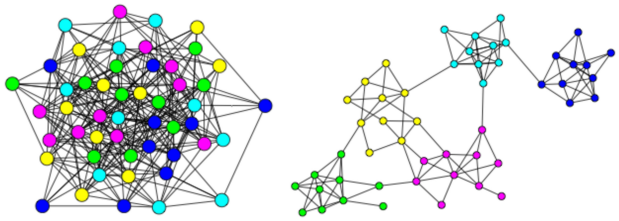


Figure 1: An instance of ER-graph (left) and planted partition graph (right). Source: Lecture note of the network analysis and modeling course in Santa Fe Institute [1].

THEOREM 3.1. *For a ER graph $G_{\text{ER}}(n,p)$ and the seed budget $k=\kappa n$ with $\kappa < \left(\frac{1-h}{2} - \frac{h}{\sqrt{\lambda}}\right)$, every $C \subset V$ with $|C|=k$ is almost surely a (γ,δ)-approximate solution as $n \to \infty$, where*

$$\delta = 1 \quad \textit{and} \quad \gamma = \begin{cases} 1+\varepsilon & \textit{for any given } \varepsilon > 0 \quad \textit{if} \quad \lambda = \omega(1), \\ 1 + \frac{2}{\frac{\sqrt{\lambda}}{2(1-h^2)}\left(\kappa - \frac{1-h}{2}\right)^2 - 1} & \textit{if} \quad \lambda = \Theta(1). \end{cases}$$

Theorem 3.1 implies that for the relatively dense and (globally) well-connected ER graph, formally for the case $\lambda = \omega(1)$, an *arbitrary* seed set C is, somewhat surprisingly, an almost optimal solution, i.e., $(\gamma,\delta) \to (1,1)$. Furthermore, we remark that in this case, the diffusion exponent Γ^* is $\lambda n \left[\kappa - \frac{1-h}{2}\right]_+^2 + o(\lambda n)$ with high probability.[1] Thus, one needs a seed budget larger than $\frac{1-h}{2}n$ in order to have an order-wise reduction in Γ^*. The near optimality of an arbitrary seeding in the dense ER graphs mainly comes from globally symmetric connectivities of nodes which makes the influencing effect by each node indistinguishable. Therefore no careful seeding mechanism is necessary for this globally well-connected graph.

3.2 Planted Partition Graphs

Second, we consider a generalized version of ER graphs and study the so-called *planted partition graph* [2], which we denote by $G_{\text{PP}}(n,p,q,\boldsymbol{\omega})$. It is a popular model, e.g., [7], for social networks with big communities (also called clusters); Given a disjoint partition of the clusters $\{V_1,...,V_m\}$, with $\bigcup_{l=1}^m V_l = V$, let the fraction of nodes in the graph that belongs to a cluster l be $\omega_l = |V_l|/n$ where $\boldsymbol{\omega} = (\omega_1,...,\omega_m) \in (0,1)^m$. For a pair of $i,j \in V$, an edge (i,j) exists between them with probability $p = \Theta(1)$ for the nodes i and j if i,j belong to a same cluster, and with probability $q < p$, otherwise. We obtain the following result, whose proof is presented in Section 4.2.

THEOREM 3.2. *For a planted partition graph $G_{\text{PP}}(n,p,q,\boldsymbol{\omega})$ and the seed budget $k=\kappa n$ with $\kappa < \frac{1-h}{2}$, every $C \subset V$ such that*

$$C \in \arg \min_{\{C':|C'|\le k\}} \max_{1\le l\le m} \left(\frac{1-h}{2}|V_l| - |C' \cap V_l|\right) \quad (8)$$

is almost surely a (γ,δ)-approximate solution as $n \to \infty$, where for any given $\varepsilon > 0$,

$$\delta = 1, \quad \textit{and} \quad \gamma = \begin{cases} 1+\varepsilon & \textit{if} \quad p/q = \omega(1), \\ 1 + \frac{2}{p\xi^2/(q+\varepsilon) - 3} & \textit{if} \quad p/q = \Theta(1), \end{cases}$$

where

$$\xi = \min_{\{\boldsymbol{\nu}\in[0,1]^m : |\boldsymbol{\nu}|_1 \le \kappa\}} \max_{1\le l\le m} \left(\frac{1-h}{2}\omega_l - \nu_l\right).$$

In particular, for the homogeneous cluster size, i.e., $\boldsymbol{\omega} = (\frac{1}{m},...,\frac{1}{m})$,

$$\xi = \frac{1}{m}\left(\frac{1-h}{2} - \kappa\right).$$

Theorem 3.2 provides a guideline on how to allocate seeds, coming from solving a "simple" min-max optimization (8) whose computational complexity is $O(1)$ (m is a given constant and only cardinality of $C' \cap V_l$ is necessary in computing the min-max solution). Intuitively the resulting seed set C in (8) allocates seeds proportionally to the size of each cluster, and intra-cluster seeding does not have to be carefully chosen. More formally, any seed set C with such an allocation is an almost optimal solution, regardless of how to seed inside each cluster if the graph is locally well-connected with big clusters whose sizes scales with respect to n and number of inter-cluster edges is ignorable comparing to intra-cluster ones, i.e., $|V_l| = O(n)$ and $p/q = \omega(1)$. Similarly to the ER graphs, we remark that in this case, a seed budget larger than $\frac{1-h}{2}n$ is required in order to have an order-wise reduction in Γ^*.We also analyze quality of the simple seeding when the inter-cluster edges are relatively substantial, i.e., $p/q = \Theta(1)$. Performance of the seeding gets closer to optimality as the ratio of intra-cluster edges to inter-cluster edges increases, i.e., higher p/q.

For locally well-connected graphs with clusters, it is necessary to intelligently exploit their clustering characteristics, where the network-wide diffusion time is governed by both *(a)* intra-cluster diffusion and *(b)* inter-cluster correlation. In locally well-connected with *big* clusters such as $G_{\text{PP}}(n,p,q,\boldsymbol{\omega})$, the intra-cluster diffusion Γ^* in each V_l dominates the inter-cluster correlation between V_l and $V_{l'}$ with $l \neq l'$. Hence it suffices to focus on how much seed budget is distributed to each (big) cluster depending on its size.

3.3 Geometrically Structured Graphs

Third, we consider locally well-connected graphs with *small* clusters. Those graphs include geometrically structured graphs such as planar and d-dimensional graphs. In these graphs, the *inter-cluster correlation* dominantly determines the network-wide diffusion speed, and hence seeds should be selected with goal of removing the correlation. Different from the earlier two types of graphs, we here take an approach that rather than studying a particular type of graph, we first propose an algorithm and then study a sufficient condition that ensures good diffusion performance and is satisfied in the well-known geometrically structured graphs such as planar and d-dimensional graphs.

One of achieving the goal of removing inter-cluster correlation would be to seed the *border nodes* among small clusters. Motivated by this, we design a generic algorithm, called PaS (Partitioning and Seeding) (see **Algorithm 1** for a formal description) for finding good seeds. As the name implies, PaS has two phases: (i) partitioning and (ii) seeding, as elaborated in what follows.

(i) Partitioning phase: In this phase, PaS finds a partitioning with, a finite number of node clusters, where the number of clusters are chosen appropriately, depending on the underlying graph topologies. Except for a special cluster, say V_0, which will be used as the initial seed set, PaS will find the seeds contained in each cluster by the seeding phase.

(ii) Seeding phase: In this phase, PaS runs in multiple rounds, where it starts from the initial seed set V_0 (step **2-1**) and the seed set C increases by one in each round, until the entire seed set size becomes the target budget k. Let G_l and C_l be the subgraph induced

[1] Here $[x]_+$ is x if $x > 0$ and 0 otherwise.

[2] This is often referred to as the stochastic block model.

and the seed contained, by l-th cluster V_l, respectively. The seeding phase consists of two sub-phases (a) partition selection and (b) seed selection. In (a), PaS finds the partition $l^\star$ that has the slowest diffusion time with the current seed set C_l (step **2-1**). In (b), for the chosen partition $l^\star$, we replace the existing seeds $C_{l^\star}$ by completely new set of seeds whose size increases by one. The new seed set is chosen such that the diffusion time in cluster $l^\star$ is minimized (step **2-2**). Finally, the temporary seed C is updated by a new seed set in cluster $l^\star$, which is repeated until $|C| = k$ (steps **2-3** and **2-4**). The choices of partition $\{V_0, V_1 \ldots, V_m\}$ in step **1** determines the performance and complexity of the PaS algorithm, where we will consider different choices for different social networks for rigorous analysis.

Now, we are ready to present the performance guarantees of the PaS algorithm. To that end, we introduce a notation: E_l is the edge set of the subgraph induced by $V_l \cup V_0$, where V_l is the l-th cluster resulting from the partitioning phase.

THEOREM 3.3. *For given graph $G = (V, E)$ and seeding budget $k = \kappa n$ with $\kappa \in (0, 1)$, suppose that $\{V_l : l = 0, 1, \ldots, m\}$ in the partitioning phase of the PaS algorithm has the following condition:*

For some $\varepsilon \in (0, 1)$,

$$|V_0| \le \varepsilon n \quad \textit{and} \quad |V_l| = O(1), \quad \textit{for all } l = 1, ..., m. \tag{9}$$

Then, the PaS algorithm outputs a $(1, 1 - \frac{\varepsilon}{\kappa})$-approximation solution and its seeding phase takes $O(n^2)$ time.

The proof of Theorem 3.3 is presented in Section 4.3. Theorem 3.3 implies that if there exists an algorithm finding a 'good' partition (i.e., $|V_0|/n \le \varepsilon$ for some small $\varepsilon > 0$) with small clusters (i.e., $V_l = O(1)$), as specified in the condition (9), the PaS algorithm outputs an almost optimal solution. Note that V_0 corresponds to the set of border nodes among clusters. This condition (9) does not always hold. However, for the following classes of social networks, polynomial-time algorithms are known for computing such a partition satisfying the condition for *any* $\varepsilon = \Omega(1)$ [22].[3]

- ***d*-dimensional Graph.** A graph is called called a d-dimensional graph, denoted by $G_{d\mathrm{D}}(n, d, D, R)$, if each node i can be embedded to a position π_i in $\mathbb{R}^d$ such that $(i, j) \in E$ implies that the Euclidean distance between π_i and π_j is less than R and any cube of volume of B contains at most $D \cdot B$ nodes, where $d, D, R = O(1)$.
- ***Planar Graph.*** A planar graph, denoted by $G_{\mathrm{PL}}(n, \Delta)$, can be drawn on the plane without intersection of edges except nodes which is endpoints of edges and its maximum degree $\Delta = O(1)$.

Therefore, we can state the following corollary of Theorem 3.3.

COROLLARY 3.1. *For a d-dimensional graph $G_{d\mathrm{D}}(n, d, D, R)$ or planar graph $G_{\mathrm{PL}}(n, \Delta)$ and seeding budget $k = \kappa n$ with $\kappa \in (0, 1)$, there exists a polynomial-time[4] algorithm such that it outputs a $(1, 1 - \varepsilon)$-approximation solution for any $\varepsilon \in (0, 1)$.*

In Section 6, we will show that PaS algorithm shows indeed a good performance for a real social graph, showing its practicability.

[3]In fact, the author [22] considers polynomially-growing graphs and minor-excluded graphs, where d-dimensional graphs and planar graphs are their special cases, respectively.

[4]It is a polynomial with respect to n, but may be exponential with respect to $1/\varepsilon$.

Input: Graph $G = (V, E)$ and seed budget k

Output: Seed set C^{PaS}

1. Partitioning phase.

Construct a partition $\{V_l : l = 0, 1, \ldots, m\}$, where there exists no edge between V_l and $V_{l'}$ for all $l \neq l' \ge 1$,

$$\bigcup_{l=0}^{m} V_l = V \quad \text{and} \quad V_l \cap V_{l'} = \emptyset, \quad \text{for all } l \neq l' \ge 0.$$

We call V_0 "separator cluster" and each component V_l becomes a cluster, i.e., $m + 1$ is the number of clusters found in this phase.

2. Seeding phase.

2-1. Seed V_0, i.e., $C \leftarrow V_0$.

2-2. <u>*Cluster selection.*</u>
Find a cluster $1 \le l^* \le m$ such that

$$l^* \in \arg \max_{1 \le l \le m : |C_l| < |V_l|} \Gamma^*(G_l, C_l \cup V_0),$$

where G_l is the subgraph induced by $V_l \cup V_0$ and C_l is the set of seeds in V_l, i.e., $C_l = C \cap V_l$.

2-3. <u>*Seed selection in the selected cluster.*</u>
Find a new seed set D in V_{l^*} such that

$$D \in \arg \min_{D' \subset V_{l^*} : |D'| = |C_{l^*}| + 1} \Gamma^* \left(G_{l^*}, D' \cup V_0 \right).$$

2-4. Update $C \leftarrow (C \setminus C_{l^*}) \cup D$, and repeat the steps **2-2**, **2.3**, and **2-4** whenever $|C| < k$.

3. Terminate. Output C.

Algorithm 1: PaS (Partitioning and Seeding) Algorithm

4. PROOFS OF THEOREMS

This section provides the proofs of Theorems 3.1, 3.2 and 3.3.

4.1 Proof of Theorem 3.1

We first present the proof of Theorem 3.1 in this section. Consider Erdős-Rényi graph $G_{\mathrm{ER}}(n, p)$ and seed budget $k = \kappa n$. We will show that for $\kappa < \left(\frac{1-h}{2} - \frac{h}{\sqrt{\lambda}} \right)$, the following event occurs almost surely as $n \to \infty$:

$$\mathcal{L} \le \frac{\Gamma^*(C)}{\lambda n} \le \mathcal{U}, \quad \text{for all } C \text{ with } |C| = k, \tag{10}$$

where

$$\mathcal{L} = \left(\kappa - \frac{1-h}{2} \right)^2 - \frac{2(1-h^2)}{\sqrt{\lambda}},$$
$$\mathcal{U} = \left(\kappa - \frac{1-h}{2} \right)^2 + \frac{2(1-h^2)}{\sqrt{\lambda}}.$$

The above inequality (10) implies that $\Gamma^*(C)$ is highly concentrated on the interval $[\mathcal{L}, \mathcal{U}]$ for any arbitrary seed set C such that $|C| = k$. Then, we should have $\gamma = \mathcal{U}/\mathcal{L}$ from Definition 3.1. Theorem 3.1 is a direct implication of (10), because when $\lambda = \lambda_n = \omega(1)$ for any given $\varepsilon > 0$, we can find sufficiently large n such that $\mathcal{U}/\mathcal{L} = 1 + \varepsilon$, and when $\lambda = \omega(1)$ we can re-express

$\mathcal{U}/\mathcal{L}$ as in Theorem 3.1. In the rest of this section, we focus on the proof of (10).

To begin with, recall the energy function $H(\boldsymbol{x})$ in (7). For convenience, we abuse the terminology and define the energy function $H(S)$ for a set $S \subset V$ (not for a state $\boldsymbol{x}$ as in (7)) as:

$$H(S) \quad = \quad \text{cut}(S, V\backslash S) - \sum_{i\in S} h|N(i)|$$

where $\text{cut}(A, B)$ is the cardinality of the set $\{(i, j) \in E \mid i \in A, j \in B\}$ for two disjoint subsets $A, B \subset V$. Note that the above definition coincides with the original definition (5) by setting $x_i = 1$ if and only if $i \in S$. Using this energy function, one can express the function $\Gamma^*(C)$ in (7) by:

$$\Gamma^*(C) = \max_{C\subset S_0\subset V} \min_{\underline{S}:S_0\to V} \max_{t<|\underline{S}|} \left[H(S_t) - H(S_0)\right], \qquad (11)$$

where for $A \subset V$, $\underline{S} : A \to V$ is a monotone sequence of sets, $A = S_0, S_1, ..., S_{|\underline{S}|} = V$ such that $S_{t-1} \subset S_t$ and $S_t\backslash S_{t-1}$ is a vertex in $V\backslash A$ for $1 \le t \le |\underline{S}|$.

To show the concentration of Γ^*, we first show the concentration of the energy function H, as stated in the next lemma whose proof is presented in Section 5.1.

LEMMA 4.1. *Consider Erdős-Rényi graph $G_{\text{ER}}(n, p)$ with $\lambda = np = \Omega(1)$. The following events occurs almost surely as $n \to \infty$:*

$$|H(S) - a(|S|)| \le \eta(|S|),$$

where

$$a(s) = (1 - h)s(n - s)p - hs(s - 1)p,$$
$$\eta(s) = (1 - h)\sqrt{2\lambda s(n - s)} + 2h\sqrt{\lambda s(s - 1)}.$$

In Lemma 4.1, $H(S)$ is bounded by $a(|S|) \pm \eta(|S|)$ which depends only cardinality of $|S|$. Thus, the paths, which are taken in $\min$ of Γ^*, have same bounds if they have same start S_0. Hence we have following:

$$\frac{\Gamma^*(C)}{\lambda n} = \frac{1}{\lambda n} \max_{C\subset S_0\subset V} \min_{\underline{S}:S_0\to V} \max_{t<|\underline{S}|} \left[H(S_t) - H(S_0)\right]$$
$$\le \frac{1}{\lambda n} \max_{|C|\le s_1\le s_2} a(s_2) + \eta(s_2) - a(s_1) + \eta(s_1) \qquad (12)$$
$$= O\left(\frac{1}{n}\right) + \max_{\kappa\le\sigma_1\le\sigma_2} \widehat{a}(\sigma_2) + \widehat{\eta}(\sigma_2) - \widehat{a}(\sigma_1) + \widehat{\eta}(\sigma_1), \qquad (13)$$

where

$$\widehat{a}(\sigma) = (1 - h)\sigma(1 - \sigma) - h\sigma^2,$$
$$\widehat{\eta}(\sigma) = \frac{1 - h}{\sqrt{\lambda}} + \frac{2h}{\sqrt{\lambda}}\sigma.$$

In (12), we have max over $|C| \le s_1 \le s_2$ since $C \subset S_0 \subset S_t$ for $t < |\underline{S}|$. Also, in (13), the $O(\frac{1}{n})$ term is from $O\left(\frac{1}{n} + \frac{1}{\lambda n}\right)$ since we have $\lambda = \Omega(1)$.

We bound a, η by $\widehat{a}, \widehat{\eta}$ for achieving an upper bound of a succinct close-form for $\frac{\Gamma^*(C)}{\lambda n}$. However, we note that one can directly consider (12) and obtain a tighter (but of a complicated form) upper bound for $\frac{\Gamma^*(C)}{\lambda n}$. Now it is not hard to check the maximum in (13) is

$$\left(\kappa - \left(\frac{1 - h}{2} - \frac{h}{\sqrt{\lambda}}\right)\right)^2 + \frac{2(1 - h^2)}{\sqrt{\lambda}}$$

at $\sigma_1 = \kappa$ and $\sigma_2 = \left(\frac{1-h}{2} + \frac{h}{\sqrt{\lambda}}\right)$ if $\kappa \le \left(\frac{1-h}{2} - \frac{h}{\sqrt{\lambda}}\right)$. This implies that $\frac{\Gamma^*(C)}{\lambda n} \le \mathcal{U}$. The proof of the lower bound $\frac{\Gamma^*(C)}{\lambda n} \ge \mathcal{L}$ can be obtained similarly. This completes the proof of (10) and hence that of Theorem 3.1.

4.2 Proof of Theorem 3.2

In this section, we present the proof of Theorem 3.2. Consider a planted partition graph $G_{\text{PP}}(n, p, q, \boldsymbol{\omega})$, and a seed set C' with budget $k < \frac{1-h}{2}n$ satisfying the conditions in Theorem 3.2. Then, to show Theorem 3.2, it suffices to show that the following events occur almost surely as $n \to \infty$:

$$\frac{\Gamma^*(C') - \Gamma^*(C^*)}{\Gamma^*(C^*)} \le \frac{2}{\frac{p}{(q+n^{-0.4})}\xi^2 - 3} \qquad (14)$$

where

$$\xi = \min_{\{\boldsymbol{\nu}\in[0,1]^m:|\boldsymbol{\nu}|_1\le\kappa\}} \max_{1\le l\le m} \left(\frac{1 - h}{2}\omega_l - \nu_l\right),$$

and C^* is an optimal seed set, i.e., $C^* \in \arg\min_{C:|C|\le k=\kappa n} \Gamma^*(C)$. This is because when $p/q = \omega(1)$, i.e., $q = o(1)$, we have $n^{-0.4}$ becomes arbitrarily small as $n \to \infty$, thus the result follows.

We first let G_l be the subgraph induced by each l-th cluster V_l, and E_l be the edges of G_l. We also let $E_0 = E \setminus \cup_{l=1}^m E_l$, which corresponds to the set of inter-cluster edges. Consider the "split graph" $G' = (V, E' = E \setminus E_0)$, i.e., G' is a graph removing the inter-cluster edges from G.

It is easy to have the following, which states that the difference of $\Gamma^\star$ between G and G' is bounded by the number of inter-cluster edges: For every $C \subset V$,

$$|\Gamma^*(G, C) - \Gamma^*(G', C)| \quad \le \quad 2|E_0|. \qquad (15)$$

To check the above, for A, B such that $A \subset B \subset V$, we calculate $H(B) - H(A)$ as below:

$$H(B) - H(A)$$
$$= (1 - h) \cdot \text{cut}(B \setminus A, V \setminus B) - (1 - 3h) \cdot \text{cut}(A, B \setminus A)$$
$$+ 2h \cdot \text{edge}(B \setminus A) \qquad (16)$$

where $\text{edge}(S)$ is number of edges among nodes in S, i.e., $\text{edge}(S) = |\{(i, j) \in E | i, j \in S\}|$. Note that in (16), three edge sets counted by cut and edge are disjoint. Thus, from removing an edge, change in value of (16) is at most $\max(1-h, |1-3h|, 2h) \le 2$ because of $0 < h < 1$. Also, we have $S_0 \subset S_t$ in the expression of Γ^* in (11). Hence we have (15) since G' is the graph where E_0 is removed from G.

Since the number of inter-cluster edges are stochastically dominated by a random variable with the binomial distribution $B\left(\frac{n(n-1)}{2}, q\right)$, we have:

$$\mathbb{P}\left[\frac{|E_0|}{n^2} \le \frac{q}{2} + \frac{1}{4}n^{-0.4}\right] \to 1 \quad \text{as } n \to \infty, \qquad (17)$$

where note that $\mathbb{E}[|E_0|] = q\frac{n(n-1)}{2}$.

We now present a key lemma for the proof of Theorem 3.2, stating that where $\Gamma^*(G', C')$ and $\min_C \Gamma^*(G', C)$ is located.

LEMMA 4.2. *For every C' satisfying the conditions in Theorem 3.2, the following holds almost surely as $n \to \infty$,*

$$\left|\frac{\Gamma^*(G', C')}{n^2} - \xi^2 p\right| \quad \le \quad \frac{1}{2}n^{-0.4}$$
$$\left|\min_{C:|C|\le k} \frac{\Gamma^*(G', C)}{n^2} - \xi^2 p\right| \quad \le \quad \frac{1}{2}n^{-0.4}. \qquad (18)$$

Now, combining (15), (17), and Lemma 4.2, leads to:

$$\left|\frac{\Gamma^*(C')}{n^2} - \xi^2 p\right| \le (q + n^{-0.4}) \qquad (19)$$

Furthermore, the following occurs almost surely as $n \to \infty$:

$$\frac{\Gamma^*(G,C) - \Gamma^*(G,C^*)}{n^2} \overset{(a)}{\leq} \frac{\Gamma^*(G,C') - \Gamma^*(G',C^*)}{n^2} + \frac{2|E_0|}{n^2}$$
$$\overset{(b)}{\leq} \frac{\Gamma^*(G',C')}{n^2} - \min_{C:|C|\leq k} \frac{\Gamma^*(G',C)}{n^2} + \frac{4|E_0|}{n^2}$$
$$\overset{(c)}{\leq} n^{-0.4} + \frac{4|E_0|}{n^2} \overset{(d)}{\leq} 2(q + n^{-0.4}), \tag{20}$$

where (a) is from (15), (b) is from (15) and the inequality: $\min_C \Gamma^*(G',C) \leq \Gamma^\star(G',C^\star)$, (c) is from Lemma 4.2, and finally (d) is from (17).

Then, (14) is a direct implication of (19) and (20), noting the the bound $\frac{\Gamma^*(C')}{\Gamma^*(C^*)} \leq \frac{\xi^2 p - 2(q+n^{-0.4})}{\xi^2 p - 3(q+n^{-0.4})}$. This completes the proof.

4.3 Proof of Theorem 3.3

This section provides the proof of Theorem 3.3. It is not hard to check the complexity of the seeding phase is $O(n^2)$ for the following reason: In the seeding phase, we have total $k = O(n)$ iterations. In each iteration, the number of clusters in the partition satisfying **P1** is $O(n)(= m)$. Further, the subphases of partition selection take $O(m)$ and $O(1)$ times, respectively, because using $|V_l| = O(1)$, $l = 1, \ldots, m$, we can compute the value Γ^* in each subgraph G_l in $O(1)$ time (note that the nodes in V_0 are already seeded).

We henceforth focus on the approximation quality of the output from the PaS algorithm. To this end, one can observe that the output C^{PaS} of the PaS algorithm minimizes Γ^* in each subgraph G_l for the budget allocation $v_l^{\text{PaS}} = |C^{\text{PaS}} \cap V_l|$, i.e.,

$$C_l^{\text{PaS}} \in \arg \min_{\{C_l \subset V_l : |C_l| \leq |C_l^{\text{PaS}}|\}} \Gamma^*(G_l, C_l \cup V_0), \tag{21}$$

where $C_l^{\text{PaS}} = C^{\text{PaS}} \cap V_l$. Recall that G_l is the subgraph induced by $V_l \cup V_0$. In addition, we use the following lemma whose proof is given in Section 5.3.

LEMMA 4.3. *For every seed set C such that $V_0 \subset C \subset V$,*

$$\Gamma^*(C) = \max_{l=1,\ldots,m} \Gamma^*(G_l, C_l \cup V_0),$$

where $C_l = C \cap V_l$.

From Lemma 4.3 and (21), we have that

$$\Gamma^*(C^{\text{PaS}}) = \max_{1\leq l\leq m} \min_{\{C_l \subset V_l : |C_l| \leq |C_l^{\text{PaS}}|\}} \Gamma^*(G_l, C_l \cup V_0). \tag{22}$$

Now we state the following key lemma, where its proof uses the above characterization of $\Gamma^*(G, C^{\text{PaS}})$ and is presented in Section 5.4.

LEMMA 4.4. *Given graph $G = (V, E)$ and budget k, the output C^{PaS} of the PaS algorithm satisfies that*

$$C^{\text{PaS}} \in \arg \min_{C:|C|\leq k, V_0 \subset C} \Gamma^*(C).$$

From Lemma 4.4, it follows that C^{PaS} is a $(1, 1-\frac{\varepsilon}{\kappa})$-approximation solution, since

$$\Gamma^*(C^{\text{PaS}}) = \min_{C:|C|\leq k, V_0\subset C} \Gamma^*(C) \leq \min_{C:|C|\leq k-|V_0|} \Gamma^*(C) \leq \min_{C:|C|\leq k(1-\frac{\varepsilon}{\kappa})} \Gamma^*(C),$$

where we use $|V_0| \leq \varepsilon n$, $k = \kappa n$ and the monotone property of Γ^*, i.e., for all A, B such that $A \subset B \subset V$, $\Gamma^*(B) \leq \Gamma^*(A)$. This completes the proof of Theorem 3.3.

5. PROOF OF LEMMAS

This section provides the proofs of Lemmas 4.1, 4.2, 4.3 and 4.4.

5.1 Proof of Lemma 4.1

Consider a subset $S \subset V$, where let $s = |S|$. For $i \in S$, we can split $N(i)$ into two disjoint sets as $N(i) = \Big(N(i) \setminus S\Big) \bigcup \Big(N(i) \cap S\Big)$. Using this separation, $H(S)$ in (11) can be written as:

$$H(S) = (1-h)\text{cut}(S, V\setminus S) - h \sum_{i\in S} |N(i) \cap S|. \tag{23}$$

In the ER graph, note that $\text{cut}(S, V \setminus S)$ and $\frac{1}{2}\sum_{i\in S} |N(i) \cap S|$ follows the binomial distributions $B(s(n-s), p)$ and $B(s(s-1)/2, p)$, respectively. Then, from the Chernoff's bound, we have

$$\mathbb{P}\left[\Big|\text{cut}(S, V\setminus S) - ps(n-s)\Big| \geq \sqrt{2\lambda s(n-s)}\right] \leq 2\exp(-n), \tag{24}$$

$$\mathbb{P}\left[\Big|\frac{1}{2}\sum_{i\in S} |N(i)\cap S| - ps(s-1)/2\Big| \geq \sqrt{\lambda s(s-1)}\right] \leq 2\exp(-n). \tag{25}$$

Thus, by applying the union bound to (24) and (25) and using (23), it follows that

$$\mathbb{P}\Big[\big|H(S) - a(s)\big| \geq \eta(s)\Big] \leq 4\exp(-n), \tag{26}$$

where $a(s)$ and $\eta(s)$ are defined in Lemma 4.1. Finally, we complete the proof using the above inequality:

$$\mathbb{P}\left[\bigcap_{S\subset V} \left[\big|H(S) - a(|S|)\big| \leq \eta(|S|)\right]\right] \geq 1 - 4\exp(-n)\cdot 2^n \to 1 \quad \text{as } n \to \infty,$$

where we use the union bound and (26) for the first inequality.

5.2 Proof of Lemma 4.2

We first note that each subgraph G_l is an ER graph $G_{\text{ER}}(\omega_l n, p)$ where its $\Gamma^*(G_l, \cdot)$ was already studied in Section 4.1. Hence, from (13) with $p = \Theta(1)$, we have $\widehat{\eta}(\sigma) = O(n^{-0.5}) = o(n^{-0.4})$.[5] Thus, for any $C_l \subset V_l$ we have almost surely as $n \to \infty$:

$$\frac{\Gamma^*(G_{\text{ER}}(\omega_l n, p), C_l)}{n^2} = \begin{cases} \left(\frac{1-h}{2}\omega_l - \nu_l\right)^2 p + \frac{1}{2}n^{-0.4} & \text{if } \nu_l \leq \frac{1-h}{2} \\ \frac{1}{2}n^{-0.4} & \text{otherwise,} \end{cases}$$

where $\nu_l = \frac{|C_l|}{n}$. Also, we note that $|V_l| = \omega_l n = \Omega(n)$. Using the above, we have that almost surely as $n \to \infty$, for every $C_l \subset V_l$,

$$\frac{\Gamma^*(G_l, C_l)}{n^2} = \left(\max\left(\frac{1-h}{2}\omega_l - \nu_l, 0\right)\right)^2 p + \frac{1}{2}n^{-0.4}. \tag{27}$$

Since G' consists of disconnected subgraphs $G_1, ..., G_m$, we provide the following which implies that the value Γ^* in the entire graph is decided by the maximum of the corresponding values in subgraphs: for every seed set $C \subset V$,

$$\Gamma^*(G', C) = \max_{l=1,\ldots,m} \Gamma^*(G_l, C_l). \tag{28}$$

[5]Here we have $\lambda = np = \Theta(n)$.

The proof of (28) is almost identical to that of Lemma 4.3, and we omit it for brevity.

Now observe that for every $C \subset V$ with $\frac{|C|}{n} \leq \kappa \leq \frac{1-h}{2}$, there exists l such that $\frac{|C_l|}{n} = \nu_l \leq \frac{1-h}{2}\omega_l$. Thus, from (27) and (28), it follows that for every $C \subset V$ such that $|C| \leq k \leq \frac{1-h}{2}n$,

$$\frac{\Gamma^*(G', C)}{n^2} = \left(\max_{1\leq l\leq m}\left(\frac{1-h}{2}\omega_l - \nu_l\right)\right)^2 p + \frac{1}{2}n^{-0.4}, \quad (29)$$

where $\nu_l = \frac{|C_l|}{n}$.

Therefore, it suffices to show the following:

$$\left|\max_{1\leq l\leq m}\left(\frac{1-h}{2}\omega_l - \nu_l'\right) - \xi\right| \leq \frac{1}{2}n^{-0.4} \quad (30)$$

where $\nu_l' = \frac{|C_l'|}{n}$.

Since we consider C' satisfying (8), $\max_{1\leq l\leq m}\left(\frac{1-h}{2}\omega_l - \nu_l'\right)$ and ξ are the same except that the $\min$ is taken over $\boldsymbol{\nu}$ consisting of continuous v_l in ξ but we have the discreteness of $\nu_l' = \frac{|C'\cap V_l|}{n}$. Due to this discreteness, ξ and $\max_{l=1,\ldots,m}\frac{f(G_l, C_l')}{n}$ have at most $\frac{1}{n}$ difference which is less than $n^{-0.4}$ as $n \to \infty$. This completes the proof.

5.3 Proof of Lemma 4.3

We use proof by induction with respect to the number of clusters, i.e. m. The following claim states formally the base case $m = 2$, where its proof is presented in Appendix B.

PROPOSITION 5.1. *For given $G = (V, E)$, consider a partition $\{V_l : l = 0, 1, 2\}$, where there exists no edge between V_1 and V_2,*

$$\bigcup_{l\in\{0,1,2\}} V_l = V \quad \textit{and} \quad V_l \cap V_{l'} = \emptyset, \quad \textit{for all } l \neq l' \geq 0.$$

Then, it follows that for any seed set C such that $V_0 \subset C \subset V$,

$$\Gamma^*(C) = \max_{l=1,2}\Gamma^*(G_l, C_l \cup V_0),$$

where $G_l = (V_l \cup V_0, E_l)$ is the induced subgraph by $V_l \cup V_0$ and $C_l = C \cap V_l$.

We now consider two subgraphs $G_1 = (V_1 \cup V_0, E_1)$ and $G_{\text{-}1} = (V_{\text{-}1} \cup V_0, E_{\text{-}1})$ where

$$V_{\text{-}1} = \cup_{l=2}^m V_l \text{ and } E_{\text{-}1} = \cup_{l=2}^m E_l.$$

Note that the separator V_0 also partitions G into G_1 and $G_{\text{-}1}$ which are the subgraphs induced by V_1 and $V_{\text{-}1}$, respectively. Then, from the construction of $G_{\text{-}1}$ and Proposition 5.1, for any seed set C such that $V_0 \subset C \subset V$, we have

$$\Gamma^*(C) = \max\left\{\Gamma^*(G_1, C_1 \cup V_0), \Gamma^*(G_{\text{-}1}, C_{\text{-}1} \cup V_0)\right\},$$

where $C_{\text{-}1} = C \cap V_{\text{-}1}$.

Observe that V_0 also partitions $G_{\text{-}1} = (V_{\text{-}1} \cup V_0, E_{\text{-}1})$ into two subgraphs $G_2 = (V_2 \cup V_0, E_2), G_{\text{-}2} = (V_{\text{-}2} \cup V_0, E_{\text{-}2})$ where $V_{\text{-}2} = \cup_{l=3}^m V_l$ and $E_{\text{-}2} = \cup_{l=3}^m E_l$. Then, one can also apply Proposition 5.1 to $G_{\text{-}1}$ again: for any seed set C such that $V_0 \subset C \subset V_{\text{-}1}$,

$$\Gamma^*(G_{\text{-}1}, C_{\text{-}1} \cup V_0) = \max\left\{\Gamma^*(G_2, C_2 \cup V_0), \Gamma^*(G_{\text{-}2}, C_{\text{-}2} \cup V_0)\right\},$$

where $C_{\text{-}2} = C \cap V_{\text{-}2}$. Thus, we have, for any seed set C such that $V_0 \subset C \subset V$,

$$\Gamma^*(C) = \max\left\{\Gamma^*(G_{\text{-}2}, C_{\text{-}2} \cup V_0), \max_{l=1,2}\Gamma^*(G_l, C_l \cup V_0)\right\}.$$

This provides the proof of Lemma 4.3 for the case $m = 3$. One can repeat this procedure to complete the proof of Lemma 4.3.

5.4 Proof of Lemma 4.4

We use proof by contradiction. To this end, suppose that there exists $C^* \neq C^{\text{PaS}}$ such that, $|C^*| = k$, $V_0 \subset C^* \subset V$ and

$$\Gamma^*(C^{\text{PaS}}) > \Gamma^*(C^*). \quad (31)$$

Let $C_l^{\text{PaS}} = C^{\text{PaS}} \cap V_l$ and $C_l^* = C^* \cap V_l$. Then, from $C^* \neq C^{\text{PaS}}$, there must exist l' such that

$$|C_{l'}^{\text{PaS}}| > |C_{l'}^*|. \quad (32)$$

The above inequality implies that the PaS algorithm selects the cluster l' (in step **2-3**) more than $|C_{l'}^*|$ times, where we say that it does for the $|C_{l'}^*| + 1$ time at the *t-th* iteration of the seeding phase. This means that at the end of the $(t-1)$*-th* iteration, the set of seeds in the cluster l' has cardinality $|C_{l'}^*|$ and the largest Γ^* among clusters, i.e., $|C_{l'}^{\text{PaS}}(t-1)| = |C_{l'}^*|$, and

$$\begin{aligned}\Gamma^*(C^{\text{PaS}}(t-1)) &= \Gamma^*(G_{l'}, C_{l'}^{\text{PaS}}(t-1)) \\ &= \min_{C_{l'}\subset V_{l'}:|C_{l'}|\leq|C_{l'}^{\text{PaS}}(t-1)|}\Gamma^*(G_{l'}, C_{l'}\cup V_0) \\ &= \min_{C_{l'}\subset V_{l'}:|C_{l'}|\leq|C_{l'}^*|}\Gamma^*(G_{l'}, C_{l'}\cup V_0) \quad (33)\end{aligned}$$

where $C^{\text{PaS}}(t-1)$ denotes the intermediate seed set at the end of the $(t-1)$*-th* iteration of the seeding phase. Therefore, it follows that

$$\begin{aligned}\Gamma^*(C^{\text{PaS}}) &\overset{(a)}{\leq} \Gamma^*(C^{\text{PaS}}(t'-1)) \\ &\overset{(b)}{=} \min_{C_{l'}\subset V_{l'}:|C_{l'}|\leq|C_{l'}^*|}\Gamma^*(G_{l'}, C_{l'}\cup V_0) \\ &\leq \Gamma^*(G_{l'}, C_{l'}^*\cup V_0) \\ &\leq \max_{1\leq l\leq m}\Gamma^*(G_l, C_l^*\cup V_0) \\ &\overset{(c)}{=} \Gamma^*(C^*),\end{aligned}$$

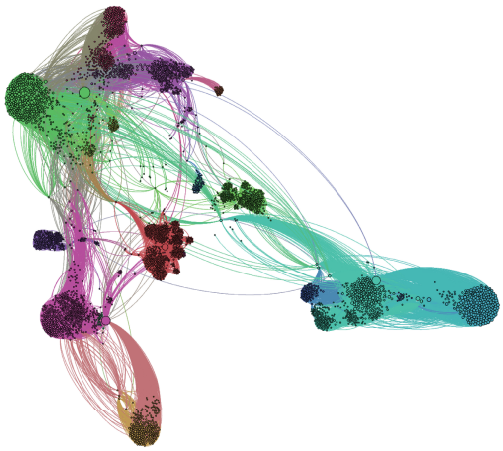
where (a) is from the fact that the PaS algorithm keeps reducing Γ^* at every iteration, (b) is due to (33), and (c) uses Lemma 4.3. This conflicts to (31), and completes the proof of Lemma 4.4.

6. SIMULATION RESULTS

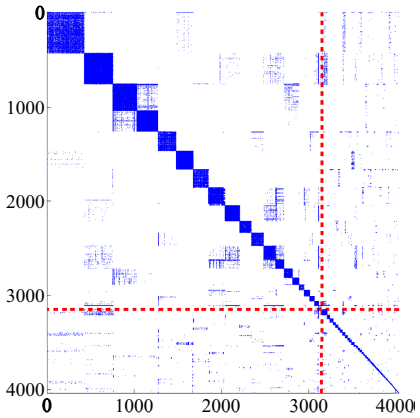
In this section, we perform simulations using a real social network graph and show how our theoretical findings can be applied to the diffusion speed maximization in practice. Guided by the implications drawn from the analytic results based on three graph classes, we propose a practical, heuristic seeding algorithm and show how it performs, compared to other seeding algorithms.

6.1 Setup

A real-world social network. We use a topology data set of the social network extracted from 10 ego networks among Facebook users originally used in [30]. The network is an undirected graph consisting of 4039 nodes and 88234 edges where each node corresponds to a Facebook account and an edge corresponds to a social relationship (called "FriendList") in Facebook. Figure 2(a) depicts a blueprint of the network, where we observe that there exist about 10 giant clusters, corresponding to 10 ego networks. Figure 2(b) presents the adjacency matrix of the graph with partitioning, showing 90 clusters that is obtained from the partitioning scheme in [41] (we use this for our seeding algorithm, as will be described shortly). We use $\beta = 10$ for the degree of rationality and vary h from 0 to 1 to investigate the impact of the difference between new and old technologies. We are interested in the regime of users are sufficiently rational and hence we tested various values of β larger than



(a) A blueprint of the Facebook ego networks consisting of 4039 users and 88234 edges.



(b) The adjacency matrix of the Facebook ego networks

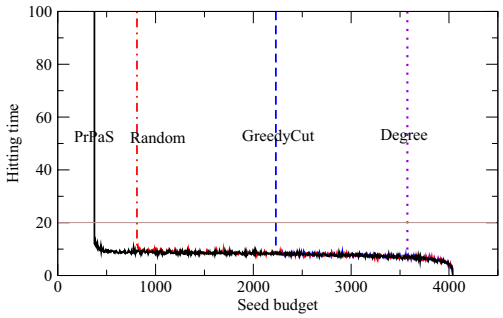
Figure 2: The Facebook ego networks used for our simulation.

10. They resulted in a similar trend and thus we just report the case of $\beta = 10$ in this paper due to space limitation.

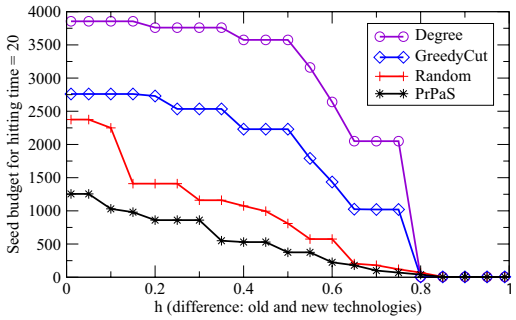
Tested seeding algorithms. We compare the performance of the following four algorithms for comparison, each of which is described in what follows.

- **Degree.** This choose k nodes in the order of their degrees.
- **GreedyCut.** This runs k iterations where at each iteration a node with the maximum number of edges is selected, and then removed from the seed candidates.
- **Random.** This selects k nodes uniformly at random.
- **PrPaS.** This first identifies the partition, say $\{V_1, ..., V_m\}$, from the given graph using the random-walk based approach [41], and then generates a seed set C whose per-cluster portion is kept equal, i.e., $|C \cap V_l|/|V_l| = k/n$ for $l = 1, ..., m$. In each cluster, seeds are selected uniformly at random.

PrPaS (Practical PaS) is the algorithm that is motivated by our theoretical findings. According to our analysis, we prefer a "good" partition consisting of locally well-connected clusters. We employ the random walk based partitioning scheme, borrowed from [41]. Then, with the resulting partition, we just balance the fraction of seeds in each cluster, so that the entire seed budget is allocated in proportion to the cluster size. This can be regarded as a practical version of **PaS** in Section 3.3 in the sense that (i) it works without explicit knowledge of h, which may be hard to be quantified in practice, and (ii) partitioning based on simple random walks is scalable and applicable to large-scale social networks. We assume the case when h is unknown, thus exact computation of $\Gamma^\star$ inside each cluster is infeasible, which is reason why we use per-cluster random seeding.



(a) Hitting times with varying seed budgets. $h = 0.5$



(b) Threshold (20) with varying h for different algorithms.

Figure 3: Simulation results for the Facebook ego networks.

6.2 Results

We compare the algorithms by the minimum seed budget with which the system hits the state $+\mathbf{1}$ in a reasonable time. For convenience, we call this minimum seed budget for a given hitting time x, *threshold* (x).

We first understand how hitting time changes with varying seed budgets. As shown in Figure 3(a), we observe that there exists a phase transition that the hitting time blows up after some seed budget, which differs across the algorithms. Due to space limitations we omit the results for other h values, where we observe a similar behavior with different seed budget leading to the hitting time blow-up. This phase transition is due to the existence of "bottleneck clusters", without which diffusion would become fast. Hence, the seeding quality can be evaluated by how efficiently such bottleneck clusters are removed by the seeding. In our setting, we see that time 20 (a horizontal line in Figure 3(a)) can be a reasonable

required hitting time to differentiate the tested algorithms. Hitting time 20 may or may not be the required time by seeders, because the absolute time should be computed by the duration of unit time and unit time can be different how actively individuals interact with each other over the given social network.

To investigate how the tested algorithms perform, we choose the time 20 as a given hitting time, and compare *threshold (20)* for all tested algorithms with varying h, whose results are shown in Figure 3(b). We first observe that across all ranges of h, **PrPaS** has the lowest threshold budget, performing significantly better than others by more than 100% difference. It is natural that for significantly high h (e.g., larger than 0.7) the performance difference is marginal because diffusion should occur very fast irrespective of the quality of seeding . **Degree** and **GreedyCut** do not perform well. Those algorithms are known to generate good seeds in epidemic-based diffusion models. This is because in epidemic-based diffusion just a contact is an efficient way of infecting neighbors, whereas in game-based models local interaction becomes much more complicated due to individuals' rational decisions. Similar phenomenon was also studied in [33] which proves that the hitting time becomes slower as individuals are connected better. **Random** significantly outperforms **Degree** and **GreedyCut**, because uniformly random seed selection allocates more seeds in larger clusters in the *average* sense. **PrPaS** performs much better than **Random** because **PrPaS** performs further optimization by considering the clustering and connectivity structure of the underlying graph.

7. CONCLUSION

In this paper, we have studied the question on how the diffusion speed of a new innovation can be maximized under a noisy game-based model, by seeding a subset of individuals (within a give budget), i.e., convincing them to pre-adopt a new innovation. By analyzing three representative graph classes, i.e., Erdős-Rényi, planted partition and geometrically structured graphs, we obtain new topological insights for the question, which does not exists in the literature for popular epidemic-based models. Our results first implies that for globally well-connected graphs, a careful seeding is not necessary. On the other hand, for locally well-connected graphs, their clustering characteristics should be understood for good seeding, where seeding inside and cross clusters are important for such graphs having big and small clusters, respectively. We believe that these new insights will provide useful tools to understand and control the sociological evolution of innovations spread over large-scale social networks.

8. REFERENCES

[1] `http://tuvalu.santafe.edu/~aaronc/courses/5352/`.

[2] R. M. Anderson and R. M. May. *Infectious Diseases of Humans*. Oxford University Press, 1991.

[3] N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Hafner Press, 1975.

[4] S. Banerjee, A. Gopalan, A. K. Das, and S. Shakkottai. Epidemic spreading with external agents. *arXiv preprint arXiv:1206.3599*, 2012.

[5] L. E. Blume. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5(3):387–424, 1993.

[6] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Infroamtion and System Security*, 10(4):13:1–13:25, 2005.

[7] K. Chaudhuri, F. C. Graham, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research-Proceedings Track*, 23:35–1, 2012.

[8] W. Chen, W. Lu, and N. Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. In *Proc. of AAAI*, 2012.

[9] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proc. of ACM SIGKDD*, 2010.

[10] J. S. Coleman, E. Katz, H. Menzel, et al. *Medical innovation: A diffusion study*. Bobbs-Merrill Company New York, NY, 1966.

[11] E. Coupechoux and M. Lelarge. Impact of clustering on diffusions and contagions in random networks. In *Proc. of IEEE NetGCooP*, 2011.

[12] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. of ACM SIGKDD*, 2001.

[13] M. Draief and A. Ganesh. A random walk model for infection on graphs: spread of epidemics rumours with mobile agents. *Discrete Event Dynamic Systems*, 21(1):41–61, Mar. 2011.

[14] G. Ellison. Learning, local action, and coordination. *Econometrica*, 61(5):1047–1071, Sep. 1993.

[15] A. Ganesh, L. Massouli, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proc. of IEEE Infocom*, 2003.

[16] A. Goyal, F. Bonchi, L. V. S. Lakshmanan, and S. Venkatasubramanian. On minimizing budget and time in influence propagation over social networks. *Social Network Analysis and Mining*, pages 1–14, 2013.

[17] H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[18] J. Hu, K. Meng, X. Chen, C. Lin, and J. Huang. Analysis of influence maximization in large-scale social networks. In *Proc. of ACM SIGMETRICS*, 2013.

[19] N. Immorlica, J. Kleinberg, M. Mahidian, and T. Wexler. The role of compatibility in the diffusion of technologies through social networks. In *Proc. of ACM EC*, 2007.

[20] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[21] Y. Jin, J. Ok, Y. Yi, and J. Shin. On the impact of global information on diffusion of innovations over social networks. In *Proc. of IEEE Infocom NetSci*, 2013.

[22] K. Jung. *Approximate inference: decomposition methods with applications to networks*. PhD thesis, Massachusetts Institute of Technology, 2009.

[23] M. Kandori, G. J. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56, Jan. 1993.

[24] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of ACM SIGKDD*, 2003.

[25] D. Kempe, J. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *Proc. of Intl. Colloq. on Automata, Languages and Programming*, pages 1127–1138. Springer, 2005.

[26] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. In *Proc. of the Royal society of London. Series A*, 1932.

[27] M. Lelarge. Coordination in network security games. In *Proc. of INFOCOM*, 2012.

[28] M. Lelarge. Diffusion and cascading behavior in random networks. *Games and Economic Behavior*, 75(2):752–775, 2012.

[29] S. Liu, L. Ying, and S. Shakkottai. Influence maximization in social networks: An ising-model-based approach. In *Proc. of IEEE Allerton*, 2010.

[30] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Proc. of NIPS*, 2012.

[31] D. McFadden. *Chapter 4: Conditional logit analysis of qualitative choice behavior in Frontiers in Econometrics*. Academic Press, New York, 1973.

[32] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

[33] A. Montanari and A. Saberi. Convergence to equilibrium in local interaction games. In *Proc. of IEEE FOCS*, 2009.

[34] D. Mookherjee and B. Sopher. Learning behavior in an experimental matching pennies game. *Games and Economic Behavior*, 7(1):62–91, 1994.

[35] S. Morris. Contagion. *The Review of Economic Studies*, 67(1):57–78, 2000.

[36] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functionsâĂŤi. *Mathematical Programming*, 14(1):265–294, 1978.

[37] P. Netrapalli and S. Sanghavi. Learning the graph of epidemic cascades. In *Proc. of ACM SIGMETRICS*, 2012.

[38] E. J. Neves and R. H. Schonmann. Behavior of droplets for a class of glauber dynamics at very low temperature. *Probability theory and related fields*, 91(3-4):331–354, 1992.

[39] E. Olivieri and M. E. Vares. *Large Deviations and Metastability*. Cambridge University Press, 2005.

[40] E. M. Rogers. *Diffusion of innovations*. Free Press, 1962.

[41] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4):1118–1123, 2008.

[42] D. Strang and S. A. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual review of sociology*, 24:265–290, 1998.

[43] H. P. Young. *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press, 1998.

APPENDIX

A. EXAMPLE IN SECTION 2.3

Consider a graph G consisting of two disconnected complete graphs G_1 and G_2, both of which has n nodes. Then, it is easy to show that for any seed set C in G_1, $\Gamma^\star(G_1, C) = 0$ (we explicitly express the parameter G_1 if needed, but omit it for the graph G for notational simplicity), when $|C| \geq \frac{1-h}{2}n$, similarly applied to G_2, also. Note that $\Gamma^\star(G_1, \emptyset) = \Gamma^\star(G_2, \emptyset)$, which corresponds to the diffusion exponent without any seeding. First, observe that $\Gamma^\star(\emptyset) = \Gamma^\star(G_1, \emptyset) = \Gamma^\star(G_2, \emptyset)$ due to symmetry and disconnectedness of G_1 and G_2. We now consider a seed set C_1 and C_2 in G_1, G_2, respectively, where $|C_1| = |C_2| = \frac{1-h}{2}n + 1$ (thus the diffusion exponent of both subgraphs is 0). Then, we have:

$$\Gamma^*(C_2) - \Gamma^*(\emptyset) > \Gamma^*(C_1 \cup C_2) - \Gamma^*(C_1),$$

since in LHS $\Gamma^*(C_2) = \Gamma^\star(\emptyset) = \Gamma^\star(G_1, \emptyset)$, and in RHS $\Gamma^*(C_1 \cup C_2) = 0$ and $\Gamma^*(C_1) = \Gamma^\star(G_2, \emptyset) > 0$. This disproves supermodularity of $\Gamma^\star(\cdot)$. Also, to disprove submodularity of $\Gamma^\star(\cdot)$, we additionally consider two seed sets C_1', C_2' in G_1, G_2, respectively, such that $C_l \cap C_l' = \emptyset$ with $|C_l| = |C_l'|$ for $l = 1, 2$ (note that we can do this for sufficiently large n since $\frac{1-h}{2} < \frac{1}{2}$). Then, we have:

$$\Gamma^*(C_1' \cup C_2') - \Gamma^*(\emptyset) < \Gamma^*(C_1' \cup C_2' \cup C_1 \cup C_2) - \Gamma^*(C_1 \cup C_2),$$

since, in the above, every term except $\Gamma^*(\emptyset)$ (which is positive) is 0. This also disproves submodularity of $\Gamma^\star(\cdot)$.

B. PROOF OF PROPOSITION 5.1

For notational convenience, we will use the following definitions: for subset $S_0 \subset V$ and monotone sequence of set $\underline{S} \in S_0 \to V$, we define

$$\Gamma(G, \underline{S}) = \max_{t \leq |\underline{S}|} [H(G, S_t) - H(G, S_0)] \tag{34}$$
$$\tilde{\Gamma}(G, S_0) = \min_{\underline{S}: S_0 \to V} \Gamma(G, \underline{S}).$$

Then, from the definition of Γ^*, we can write

$$\Gamma^*(G, C) = \max_{C \subset S_0 \subset V} \tilde{\Gamma}(G, S_0) \quad = \max_{C \subset S_0 \subset V} \min_{\underline{S}: S_0 \to V} \Gamma(G, \underline{S}).$$

With the given partition, the following simple equality can be derived using (23) for any subset S such that $V_0 \subset S$,

$$H(S) = H(G_1, S \cap W_1) + H(G_2, S \cap W_2). \tag{35}$$

where we let $W_1 = V_1 \cup V_0$ and $W_2 = V_2 \cup V_0$.

Let C denote a seed set such that $V_0 \subset C \subset V$. Also, let $C_1 = C \cap V_1$ and $C_2 = C \cap V_2$. To complete the proof of this proposition, we will show that the followings hold:

$$\max_{l=1,2} \Gamma^*(G_l, C_l \cup V_0) \ \leq \ \Gamma^*(C), \tag{36}$$
$$\Gamma^*(C) \ \leq \ \max_{l=1,2} \Gamma^*(G_l, C_l \cup V_0). \tag{37}$$

Proof of (36)**.** For a subset $X \subset V$ such that $C \subset X$, define

$$\mathcal{P}_l(X) \ = \ \{\underline{S}' : X \cap W_l \to W_l\},$$
$$\mathcal{Q}_l(X) \ = \ \{\underline{S} : X \cup W_l \to V\}.$$

Then, we have

$$\begin{aligned}
&\tilde{\Gamma}(X \cup W_2) \\
&= \min_{\underline{S} \in \mathcal{Q}_2(X)} \max_{t \leq |\underline{S}|} [(H(S_t) - H(S_0))] \\
&= \min_{\underline{S} \in \mathcal{Q}_2(X)} \max_{t \leq |\underline{S}|} [(H(G_1, S_t \cap W_1) + H(G_2, S_t \cap W_2))] \\
&\quad - [H(G_1, S_0 \cap W_1) + H(G_2, S_0 \cap W_2))] \qquad (\because (35)) \\
&\overset{(a)}{=} \min_{\underline{S} \in \mathcal{Q}_2(X)} \max_{t \leq |\underline{S}|} [H(G_1, S_t \cap W_1) - H(G_1, S_0 \cap W_1)] \\
&\overset{(b)}{=} \min_{\underline{S}' \in \mathcal{P}_1(X)} \max_{t \leq |\underline{S}'|} [H(G_1, S_t') - H(G_1, S_0')] \\
&= \tilde{\Gamma}(G_1, X \cap W_1)
\end{aligned} \tag{38}$$

In the above, (a) holds since $H(G_2, S_t \cap W_2) = H(G_2, W_2)$ for all t, which comes from the fact that $V_2 \cup V_0 \subset S_t$. (b) holds since there is a one-to-one correspondence between $\mathcal{P}_1(X)$ and $\mathcal{Q}_2(X)$; i.e., $\underline{S}'$ can be induced from $\underline{S}$ by $\underline{S}' = (S_0 - V_2, .., S_t - V_2, ..., V - V_2 (= W_1))$ and vice versa. Similarly, one can show that

$$\tilde{\Gamma}(X \cup W_1) = \tilde{\Gamma}(G_2, X \cap W_2). \tag{39}$$

Since $C \subset X \subset V$, it follows that

$$\begin{aligned}\Gamma^*(C) &= \max_{C \subset S_0 \subset V} \tilde{\Gamma}(S_0) \\ &\geq \max_{l=1,2} \tilde{\Gamma}(X \cup W_l) \;=\; \max_{l=1,2} \tilde{\Gamma}(G_l, X \cap W_l)\end{aligned}$$

where the last equality holds from (38) and (39).

Now by taking the maximum of $\max_{l=1,2} \tilde{\Gamma}(G_l, X \cap W_l)$ over all X such that $C \subset X \subset V$, we conclude that

$$\begin{aligned}\Gamma^*(C) \;&\geq\; \max_{C \subset X \subset V} \max_{l=1,2} \tilde{\Gamma}(G_l, X \cap W_l) \\ &=\; \max_{l=1,2} \max_{C \subset X \subset V} \tilde{\Gamma}(G_l, X \cap W_l) \\ &=\; \max_{l=1,2} \Gamma^*(G_l, C_l \cup V_0).\end{aligned}$$

This completes the proof of (36).

Proof of (37)**.** Let S_0^* and $\underline{S}^*$ be an optimal subset of V and an optimal monotone sequence of sets for G, i.e., $C \subset S_0^* \subset V$, $\underline{S}^* : S_0^* \to V$, and

$$\Gamma^*(C) = \tilde{\Gamma}(S_0^*) = \Gamma(\underline{S}^*).$$

In addition, let $\underline{S}^1 : S_0^* \cap W_1 \to W_1$ and $\underline{S}^2 : S_0^* \cap W_2 \to W_2$ be an optimal monotone sequences of sets for G_1, G_2, respectively. Then we have

$$\begin{aligned}\Gamma^*(G_1, C_1) \;&=\; \tilde{\Gamma}(G_1, S_0^* \cap W_1) \;=\; \Gamma(G_1, \underline{S}^1), \\ \Gamma^*(G_2, C_2) \;&=\; \tilde{\Gamma}(G_2, S_0^* \cap W_2) \;=\; \Gamma(G_2, \underline{S}^2).\end{aligned}$$

Now, construct $\underline{S}^1 \cup S_0^* : S_0^* \to S_0^* \cup V_1$ and $\underline{S}^1 \cup S_0^* : S_0^* \cup V_1 \to V$ such that

$$\begin{aligned}\underline{S}^1 \cup S_0^* \;&=\; (S_0^1 \cup S_0^* ..., S_t^1 \cup S_0^*, ... S_{|\underline{S}^1|}^1 \cup S_0^*), \\ \underline{S}^2 \cup V_1 \;&=\; (S_0^2 \cup V_1 ..., S_t^2 \cup V_1, ... S_{|\underline{S}^2|}^2 \cup V_1).\end{aligned}$$

Since the end of $\underline{S}^1 \cup S_0^*$ and the start of $\underline{S}^1 \cup S_0^*$ are the same (note that $S_0^1 \cup S_0^* = S_0^*$, $S_{|\underline{S}^1|}^1 \cup S_0^* = S_0^* \cup V_1 = S_0^2 \cup V_1$ and $S_{|\underline{S}^2|}^2 \cup V_1 = W_2 \cup V_1 = V$.). and $V_0 \subset S_0^*$, we can construct a new monotone sequence of sets $\underline{T} : S_0^* \to V$ by concatenating $\underline{S}^1 \cup S_0^*$ and $\underline{S}^2 \cup V_1$:

$$\begin{aligned}\underline{T} = (&S_0^*, S_1^1 \cup S_0^*, S_2^1 \cup S_0^*, ..., S_{|\underline{S}^1|}^1 \cup S_0^*, \\ &S_1^2 \cup V_1, S_2^2 \cup V_1, ..., S_{|\underline{S}^2|-1}^2 \cup V_1, V).\end{aligned}$$

Thus, we have

$$\begin{aligned}\Gamma(\underline{T}) = \max \Big(&\max_{t \leq |\underline{S}^1|} H(S_t^1 \cup S_0^*), \max_{t \leq |\underline{S}^2|} H(S_t^2 \cup V_1) \Big) \\ &- H(S_0^*).\end{aligned}$$

Using the construction of $\underline{T}$ with (34) and (35), it is not hard to check that

$$\max_{t \leq |\underline{S}^1|} H(S_t^1 \cup S_0^*) = \Gamma(G_1, \underline{S}^1) + H(S_0^*) \tag{40}$$

$$\begin{aligned}&\max_{t \leq |\underline{S}^2|} H(S_t^2 \cup V_1) = \\ &\qquad \Gamma(G_2, \underline{S}^2) + H(G_1, W_1) + H(G_2, S_0^* \cap W_2).\end{aligned} \tag{41}$$

Furthermore, using (40), (41) and (35), we have

$$\max_{t \leq |\underline{S}^1|} H(S_t^1 \cup S_0^*) - H(S_0^*) = \Gamma(G_1, \underline{S}_1^1) \tag{42}$$

$$\begin{aligned}&\max_{t \leq |\underline{S}^2|} H(S_t^2 \cup V_1) - H(S_0^*) \\ &\qquad = \Gamma(G_2, \underline{S}^2) + H(G_1, W_1) - H(G_1, S_0^* \cap W_1).\end{aligned} \tag{43}$$

Recall that the state that all players choose $+1$ has the minimum of $H(\cdot)$. Hence on the subgraph G_1, $H(G_1, \cdot)$ has the minimum at W_1, i.e., $H(G_1, W_1) = \min_{S \subset W_1} H(G_1, S)$. Thus, we have

$$H(G_1, W_1) - H(G_1, S_0^* \cap W_1) \;<\; 0.$$

Combining (42) and (43) leads us to:

$$\begin{aligned}\Gamma(\underline{T}) \;&\leq\; \max(\Gamma(G_1, \underline{S}^1), \Gamma(G_2, \underline{S}^2)) \\ &=\; \max(\tilde{\Gamma}(G_1, S_0^* \cap W_1), \tilde{\Gamma}(G_2, S_0^* \cap W_2)) \\ &\leq\; \max(\Gamma^*(G_1, C_1 \cup V_0), \Gamma^*(G_2, C_2 \cup V_0)),\end{aligned}$$

where the last inequality is due to $C \subset S_0^*$. Since $\underline{T}$ and $\underline{S}^*$ are monotone sequences of sets from $S_0^* \to V$, we have

$$\Gamma(\underline{S}^*) \;\leq\; \Gamma(\underline{T}) \;\leq\; \max_{l=1,2} \Gamma^*(G_l, C_l \cup V_0),$$

where the first equality holds by the definition of $\underline{S}^*$. This completes the proof of (37) and hence completes the proof of Proposition 5.1.