Network Intervention for Mental Disorders with Minimum Small Dense Subgroups

Bay-Yuan Hsu, Chih-Ya Shen, and Xifeng Yan

Abstract—According to the literature in psychology, the existence of small dense subgroups is closely related to many mental illnesses, such as depression, bullying, and psychotic disorders. Here, small dense subgroups refer to the small groups in the social network in which members are socially dense but have no or few links to other individuals outside the group. Therefore, in this article, we make the first attempt to address the issue of small dense subgroups with the concept of *network intervention* from Psychology. We first introduce the new notion of Δ-Subgroups (Δ-SGs) to quantify the small dense subgroups. Then, following the concept of network intervention, we formulate a new research problem, *Small Subgroup Maximum Reduction Problem (SSMP)*, to reduce the number of small dense subgroups (i.e., Δ-SGs) in the social network. We prove that SSMP is NP-Hard and propose a linear-time algorithm, namely *3-SMMTG*, to find the optimal solution for a special case of SSMP with $\Delta = 3$. We then devise a $\frac{1}{2}(1-\frac{1}{e})$-approximation algorithm, namely *ESGR*, for the general SSMP and enhance its efficiency with effective pruning methods. We conduct a 8-week evaluation study with 812 participants to validate the proposed SSMP and ESGR. The results show that the participants with the network intervention recommended by ESGR have significant improvements on Internet addiction and depression, as compared to those individuals without any intervention. We also perform experiments on 7 real datasets, and the experimental results manifest that the proposed algorithms outperform the other baselines in both efficiency and solution quality.

Index Terms—Small dense subgroups, algorithm design, social networks, mental disorders

1 Introduction

Extracting socially dense subgraphs from social networks has been extensively studied and has many practical applications [1], [2], [3], [4], [5]. However, the existence of *small dense subgroups*, i.e., small socially dense groups where group members are tightly connected within the group, but they have no or very few links to other individuals outside the group, may not always be beneficial for the members, according to the literature in psychology and sociology [6], [7], [8], [9]. Moreover, being in small dense subgroups is one important factor to many mental illnesses [10], [11].

On the other hand, *network intervention* is widely adopted in psychology to strengthen and augment individuals' social networks for desired behavior changes [12], [13], which usually involves the addition of new members and new friendships in the social network. By doing so, some small dense subgroups may become a larger dense group, and individuals may benefit from various interactions and social supports from the others. For example, network intervention is applied to the students to encourage them to build healthy social ties with others [14] for suicide prevention, and network intervention is performed to improve the student's learning [15] and improve the social communication skills of students with autism spectrum disorders [16].

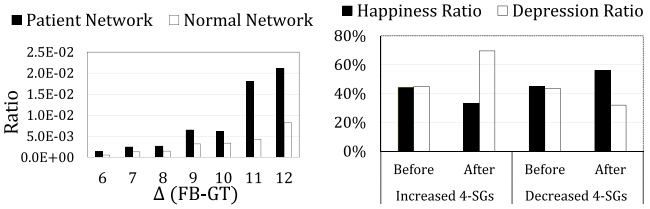
To enable network intervention to improve the individuals' well-being, certain social network property that is related to *small dense subgroups* should be modified with the inclusion of new members and friendships. An important question arises: what is the property we would like to modify by employing network intervention? Conventional measurement such as vertex degree, closeness centrality and betweenness centrality do not work well, because these measures cannot well capture the idea of small socially dense groups, where group members are tightly connected within the group, but have no or very few links to other individuals outside the group. That is, the measures mentioned above work fine to describe how dense a group is, but they fail to describe how sparse a group links to other individuals outside the group. Therefore, in this paper, we propose the new notion of Δ-Subgroups (Δ-SGs) to capture the concept of small dense subgroups in social networks, where Δ-SG is employed as the basis for quantifying small dense subgroups in social networks. Given the social network graph $G = (V, E)$, three vertices u, v, w in V form a triangle if all three edges (u, v), (u, w), (v, w) exist in E. A triangle is a Δ-SG if the degree of each of its vertices is at most Δ on G. Specifically, Δ-SG is defined as follows.

Definition 1 (Δ-SG). *Given a non-negative integer Δ and vertices $u, v, w \in V$, $\{u, v, w\}$ is a Δ-Subgroup (Δ-SG) if i) $\{u, v, w\}$ is a triangle, and ii) $deg_G(u) \leq \Delta$, $deg_G(v) \leq \Delta$, and $deg_G(w) \leq \Delta$ hold, where $deg_G(u)$ denotes the degree of vertex u on G.*

In the above definition, triangle is used as the basic element for Δ-SG because triangles are widely used to model

- *B.-Y. Hsu and X. Yan are with the Department of Computer Science, University of California, Santa Barbara, CA 93106 USA. E-mail: {soulhsu, xyan}@cs.ucsb.edu.*
- *C.-Y. Shen is with the Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan. E-mail: chihya@cs.nthu.edu.tw.*

Manuscript received 17 Dec. 2018; revised 4 Sept. 2019; accepted 2 Oct. 2019. Date of publication 25 Oct. 2019; date of current version 1 Apr. 2021. (Corresponding author: Chih-Ya Shen.) Recommended for acceptance by Z. Cai. Digital Object Identifier no. 10.1109/TKDE.2019.2949294



(a) Δ-SGs in FB-GT (b) Δ-SGs in TYP

Fig. 1. Data analysis results.

the social tightness, e.g., clustering coefficient, transitivity ratio, and k-truss [2] all employ triangles as basic elements. Recall that a small dense subgroup refers to a small group that is socially tight within but has no or very few links to other individuals, the Δ-SG serves well as a basic unit for measuring small dense subgroups because a triangle is a small complete graph of size 3 (socially dense within the small group). On the other hand, the degree upper bound, Δ, limits the number of links outward. Therefore, in this paper, we employ Δ-SGs to quantify small dense subgroups an individual participates in. Please note that Δ can be set as different values (suggested by mental health professionals) to fit various application scenarios and different social networks.

Data Analysis for Δ-SGs. To illustrate the possible correlations of Δ-SGs with mental well-being, we conduct an analysis on two real datasets: i) Facebook dataset (FB-GT) with 1,432 individuals, and ii) Taiwan Youth Project (TYP) dataset with 2,844 students,[1] as shown in Fig. 1. In FB-GT, each user is examined by mental health professionals and is labeled with an Internet addiction (IAD) score between 0 to 100. A larger score indicates a more severe Internet addiction.[2] In Fig. 1a, *Patient Network* and *Normal Network* represent the numbers of Δ-SGs the IAD patients and the normal individuals participate in, respectively. The *ratio* (y-axis) is the number of Δ-SGs (of each category) divided by the square of the network size of FB-GT. Fig. 1a shows that the *Patient Network* has a much higher Δ-SG ratio as compared to the *Normal Network*, implying the possible correlation of Δ-SGs with Internet addiction.

Further, we illustrate the impact of Δ-SGs with $\Delta = 4$ in the TYP dataset in Fig. 1b, where each of the 2,844 students in TYP is labeled with mental status scores for three consecutive years (grades 7 to 9). The screened mental status include depression (measured with standard psychological questionnaires [17]) and happiness (obtained with a 4-point Likert scale). We plot the ratios of students who report to be happy or very happy as *Happiness Ratio* and the ratio of students who have scores above mild depression as *Depression Ratio* in Fig. 1b. Please note that a higher *Happiness Ratio* or a lower *Depression Ratio* indicates a better result, i.e., more students are happy, or fewer students are suffering from depression, respectively.

1. This dataset is derived from a longitudinal panel study entitled *The Taiwan Youth Project (TYP)* with eight annual surveys, conducted by the Institute of Sociology, Academia Sinica, Taiwan. This dataset is publicly accessible at: http://www.typ.sinica.edu.tw/E/?q=node/6.

2. Our results are statistically significant with a p-value of less than 0.01.

In Fig. 1b, after the number of the Δ-SGs each student participates in increases, *Happiness Ratio* decreases and *Depression Ratio* increases significantly, indicating that more students are unhappy and suffering from depression. In contrast, after the number of Δ-SGs decreases, more students report to be happy and fewer suffer from depression. The above analysis manifests that the number of Δ-SGs each individual participates in is closely related to mental well-being. In Section 6, we conduct an evaluation study with 812 participants to validate that a network intervention that reduces the number of Δ-SGs may improve the individuals' mental health status.

To reduce the negative repercussions of small dense subgroups, *network intervention* is widely adopted in Psychology to reduce the number of small dense subgroups in the network [12], [13]. Network intervention for reducing small dense subgroups is usually carried out with the addition of new members (mental health professionals) and new links (from the mental health professionals to patients) to the network, such that multiple small dense subgroups become a large dense group. In this way, more people in the large group are linked together and can provide social support to help the mental disorder patients [10], [18].

Currently, the network intervention is performed manually. However, given a number of new members to be added, manually deciding the target patients to create links for reducing the number of small dense groups is very time-consuming and error-prone because there is an overwhelmingly large number of combinations. Moreover, simply assigning therapists to the large dense subgroups may not work well for mental health treatment because research shows that small dense subgroup, instead of large ones, seriously affects individuals' mental status [12], [13]. Therefore, in this paper, we formulate a new research problem, namely *Small Subgroup Maximum Reduction Problem (SSMP)*, to address this important issue for network intervention. The proposed SSMP aims to intervene a network G by adding at most s new vertices (mental health professionals) and at most b edges (social links) to reduce the largest number of Δ-SGs in G. With SSMP, mental health professionals are able to help the patients more effectively and efficiently.

The importance of the proposed new research problem is three-fold. i) As small dense subgroups may have negative impacts on individuals' mental well-being, we propose to quantify how an individual is associated with small dense subgroups with a new notion, Δ-SGs. ii) We analyze two real datasets to illustrate the correlation between Δ-SGs and the mental status of the individuals. Our evaluation study (in Section 6) also show that reducing the number of Δ-SGs is positively correlated to an individual's well-being. iii) In order to improve the mental well-being of the individuals, network intervention is the current practice adopted by mental health professionals to modify individuals' social network by including new members and new friendships. However, as the number of individuals is large, manually deciding the number of members and friendships to include into the social network is very time-consuming and overwhelming for mental health professionals. Therefore, a system that can automatically recommend how to intervene the individuals' social network is very helpful to the mental health professionals. Therefore, in this paper, we propose a

new research problem, named *Small Subgroup Maximum Reduction Problem (SSMP)*, to address this important need for network intervention.

Efficiently processing SSMP is very challenging because we need to examine different combinations of vertices to reduce the maximum number of Δ-SGs while adding at most b edges and s vertices. In fact, as will be proved in Section 2, SSMP is NP-Hard. Although the general SSMP is NP-Hard, we observe that a special case of SSMP with $\Delta = 3$ is still tractable. Therefore, we propose a linear-time algorithm, namely *3-SMMTG*, to obtain the optimal solution to SSMP with $\Delta = 3$. We then apply our insights in 3-SMMTG to the general case and propose a $\frac{1}{2}(1-\frac{1}{e})$-approximation algorithm for the general SSMP.

The contributions are summarized as follows.

- Based on Psychology, we propose the notion of Δ-SG to quantify small dense subgroups. We formulate a new research problem, *Small Subgroup Maximum Reduction Problem (SSMP)* to intervene the social network of patients to reduce the number of Δ-SGs in the network. To our best knowledge, this is the first work that studies the network intervention in an algorithmic aspect.
- We analyze the NP-Hardness of SSMP and propose a $\frac{1}{2}(1-\frac{1}{e})$-approximation algorithm, i.e., *ESGR*, for the general SSMP. We also propose a linear-time algorithm, namely *3-SMMTG*, to find the optimal solution for SSMP with $\Delta = 3$.
- We conduct a 8-week evaluation study with 812 participants to validate the proposed SSMP and ESGR. The results show that the participants with the network intervention recommended by ESGR have significant improvements on Internet addiction and depression, as compared to those individuals without any intervention.
- We also perform extensive experiments on 7 real datasets to evaluate the proposed algorithms. The results show that our proposed algorithms outperform the other baselines in terms of solution quality and efficiency.

The paper is organized as follows. Section 2 formulates the problem and analyzes the hardness. Section 3 reviews the works relevant to this paper. Section 4 proposes algorithm 3-SMMTG for SSMP with $\Delta = 3$. Section 5 proposes algorithm ESGR and analyzes its approximation ratio. Section 6 details the evaluation study, and Section 7 presents the experimental results. Section 8 concludes this paper.

2 Problem Formulation and Analysis

In this paper, we consider a social network of individuals $G = (V, E)$, where V is the set of individuals, and E is the set of edges representing their friendships, i.e., an edge (u, v) exists if two individuals u and v are friends.

With the definition of Δ-SG in hand (defined in Definition 1), we define $\delta(U, Y)$ as the number of Δ-SGs in the subgroup induced by the vertex subset $U \subseteq V$ and edge subset $Y \subseteq E$. Moreover, for a triangle $t = \{u, v, w\}$, we say t *covers* each of the vertices u, v, and w. Since a Δ-SG is inherently a triangle, when a Δ-SG covers a vertex u, it implies that u is one of the three vertices forming the underlying triangle of the Δ-SG. Given two integers s (called *seed constraint*) and b (called *budget constraint*), in this paper, we formulate and study the new research problem of maximizing the reduction of Δ-SGs in G by adding at most s vertices and b edges into G. The vertices can be viewed as mental health professionals in Psychology, and the edges added into G are regarded as the new friendships created by the mental health professionals to intervene the network. Please note that we only focus on the connections between the mental health professionals and the patients in our problem formulation. The new connections between patients may also be good for the treatment. However, in a real therapy, the new friendships between patients are difficult to control, because the mental health professionals cannot force the patients to build friendships with other patients.

Specifically, let A and F denote the set of vertices and edges added to G, and let $G' = (V \cup A, E \cup F)$ denote the intervened graph after including A and F, the research problem is formulated as follows.

Problem: Small Subgroup Maximum Reduction Problem (SSMP).

Given: Graph $G = (V, E)$, integers b, s, and Δ.

Find: To find an intervened simple graph $G' = (V \cup A, E \cup F)$, where A and F are the sets of new vertices and edges added to G, such that i) $G \subseteq G'$, i.e., the individuals and friendships existed in G must remain in G', ii) $|A| \leq s$, i.e., at most s vertices are added into V, iii) $|F| \leq b$, i.e., at most b edges are created, and iv) $u \in V$ and $v \in A$ hold, $\forall (u, v) \in F$, i.e., for any new edge $(u, v) \in F$, one endpoint must be an individual originally existed in V, and the other endpoint must be a newly added vertex in A.

Objective: To maximize the reduction of Δ-SGs. That is, to maximize $\delta(V, E) - \delta(V, E \cup F)$.

Note that the objective function is to maximize $\delta(V, E) - \delta(V, E \cup F)$ instead of $\delta(V, E) - \delta(V \cup A, E \cup F)$. This is because the Δ-SGs brought by adding the mental health professionals do not result in the negative repercussions. For brevity, in the following, we denote $\rho(V, E, F) = \delta(V, E) - \delta(V, E \cup F)$ so the objective function of the SSMP problem is to maximize $\rho(V, E, F)$.

In the problem formulation of SSMP, s is the maximum number of new vertices that can be added to the social network, i.e., at most s mental health professional can join the social network to create new friendships with the individuals, while b is the maximum number of new edges (the new friendship links created by connecting a mental health professional to an individual) that can be added to intervene the social network.

We incorporate these two parameters into the problem formulation of SSMP is to allow a more flexible configuration of the network intervention. In real-world intervention scenarios, the number of mental health professionals that can participate in the intervention may vary, due to their availability, expertise, and the subjects' properties. Therefore, the parameter s is included to consider the limited number of mental health professionals that can join the network intervention.

On the other hand, the mental health professionals may have limited time and resource to intervene the network (i.e., to build friendships with the individuals in the

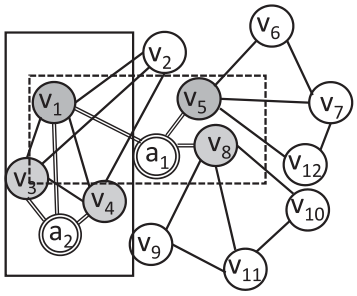


Fig. 2. Illustrative example.

network), and may not be able to build new friendships with a very large number of individuals. Therefore, the problem formulation of SSMP includes the parameter b to set an upperbound on the number of edges that can be added into the network.

Efficiently processing SSMP is very challenging because we need to consider different combinations of vertices to maximize $\rho(V, E, F)$ while satisfying the budget constraint b and the seed constraint s simultaneously. The following example shows that a trivial approach cannot obtain a good solution. Given a social network $G = (V, E)$ in Fig. 2, where $V = \{v_1, \ldots, v_{12}\}$ represents the patients (vertices a_1 and a_2 are not patients, i.e., not included in V), and the subgraph induced by $v_i, i \in [1, 12]$ has 8 3-SGs, i.e., $\delta(V, E) = 8$. Assume that a mental health professional would like to add herself to intervene the network by building at most 3 links from herself in order to break the small dense subgroups, i.e., solving an SSMP instance with $\Delta = 3$, $s = 1$ and $b = 3$. One simple approach is to greedily select 3 patients who participate in the maximum numbers of 3-SGs (i.e., v_1, v_3, v_4) and link them with the mental health professional (denoted as a_2), as shown in the left rectangle in Fig. 2. However, the patients (except the mental health professional) in the intervened network still have 4 3-SGs, and $\rho(V, E, F) = \delta(V, E) - \delta(V, E \cup F) = 8 - 4 = 4$ in this case. A better approach is to link $\{v_1, v_5, v_8\}$ to the professional (denoted as a_1 here), as shown in the dashed rectangle in Fig. 2. This intervention is much better, i.e., only 1 3-SG left for all the patients. In other words, $\rho(V, E, F) = 8 - 1 = 7$, which is also the optimal solution to this SSMP instance. In the following, we analyze the hardness of the proposed SSMP problem.

We prove the NP-Hardness of SSMP with the reduction from the *Zero Δ-SG Edge Intervention (ZEI)* problem, which is NP-Complete. Specifically, the ZEI problem is formulated as follows.

Problem: Zero Δ-SG Edge Intervention (ZEI).

Given: Graph $G_R = (V_R, E_R)$, an integer Δ, a new vertex $a \notin V$, and an integer k.

Decide: To decide whether there exists a set of k vertices $K = \{v_1, .., v_k\}$ in V_R, such that after G_R is intervened by the set F of k new edges, i.e., $F = \{(a, v_i) | \forall v_i \in K\}$, $\delta(V_R, E_R \cup F) = 0$ holds.

Please note that $\delta(V_R, E_R \cup F) = 0$ implies that the set of vertices K cover all the Δ-SGs in G_R (i.e., each Δ-SG must include at least one $v_i \in K$). This is because if there exists a Δ-SG not including any vertex in K, this Δ-SG must remain in the intervened graph $(V_R \cup \{a\}, E_R \cup F)$, and thus $\delta(V_R, E_R \cup F)$ must be nonzero.

In the following, we first prove that the above ZEI problem is NP-Complete in Theorem 1. Then, by leveraging the hardness result of ZEI, we further prove that the proposed SSMP problem is NP-Hard in Theorem 2.

Theorem 1. *ZEI is NP-Complete.*

Proof. We prove this theorem in the Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2019.2949294. □

Theorem 2. *SSMP is NP-Hard.*

Proof. Now, we turn our attention to the NP-Hardness of SSMP. Given an instance of ZEI with input graph $G_R = (V_R, E_R)$, $\widehat{\Delta}$, new vertex $a \notin V_R$, and an integer k, we create an instance of SSMP with input graph $G = (V, E)$ by setting $V = V_R$, $E = E_R$, $s = 1$, $\Delta = \widehat{\Delta}$, and budget $b = k$. We first prove the necessary condition. If G_R contains k vertices such that $\delta(V_R, E_R \cup F)$ equals zero, there must exist a group with the same set of vertices that satisfies the budget constraint b for the SSMP instance. We then prove the sufficient condition. If i) SSMP contains a group of k vertices H and ii) after we connect all the vertices in H to the new vertex in A ($|A| = 1$ due to $s = 1$) to form the new edge set F, the intervened graph $(V \cup A, E \cup F)$ has $\delta(V, E \cup F) = 0$, then ZEI must contain a solution with size exactly k as well. This proves the sufficient condition, and the theorem follows. □

3 Related Works

According to the literature in Psychology, the existence of *small dense subgroups* should be avoided in some cases [19], [20]. The negative repercussions of the patients forming small dense subgroups are also discussed extensively [6], [7], [8], [9]. For example, the existence of small dense subgroups deteriorates the effectiveness of group therapy [19]. Moreover, research indicates that if children form a small dense subgroup in the class, they are more likely to bully other children who do not belong to their small dense subgroup. Reports also indicate that in the small dense network, people with psychosis frequently experience difficulties in developing and maintaining social relationships [21]. Therefore, to help reduce the number of small dense groups in the social network, in this paper, we study the network intervention from the algorithmic aspects to effectively minimize the number of small dense subgroups.

Extracting dense subgraphs or communities from social networks has been studied extensively. Different density measures have been proposed, such as diameter [2], density [3], and clique with its variations [4]. In addition, extracting densely connected communities and overlapping communities from social networks [1], [22] and identifying dense subgraphs while satisfying requirements in different dimensions [23], [24] have also been studied extensively. For example, BigClam (Cluster Affiliation Model for Big Networks) [1] finds overlapping communities from social networks. Although the above research covers various applications, they focus mainly on extracting dense groups from social networks with different size. In contrast, to address the negative repercussions brought by small dense subgroups, this paper explores a new

research problem that aims to intervene the social network to reduce the number of small dense subgroups. Therefore, the algorithms in previous works cannot be applied to solve the proposed SSMP here.

Lin et al. [25] recruit a set of 698 students to analyze the correlations between social network structures and their health wellness states, e.g., heart rate, stress, happiness, positive attitude, and self-assessed health. Based on the analysis, the authors extract a set of features, and demonstrate that by employing machine learning techniques, they are able to predict the health behavior and wellness states of the users with high accuracy. Dhand et al. [26] analyze the personal ego network of the neurology patients and discuss the relationship between the patient's ego network and their illnesses. By analyzing users' tweets, including users' textual, visual, social attributes, and social interaction data, Lin et al. [27] propose a hybrid model that integrates the factor graph model with Convolution Neural Network to detect a user's stress level. Also, Fraga et al. [28] analyze the user activities and interactions on Reddit to discuss the patterns of the posts on Reddit communities related to mental disorders. They observe that for the Reddit communities, i.e., Depression, SuicideWatch, Anxiety, and Bipolar, the interaction patterns are similar. Moreover, larger discussion trees are usually formed by the posts asking for help. For eating disorder, Wang et al. [29] analyze the interaction patterns of the communication network on Twitter of a set of self-identified eating disorder patients. They discover two eating disorder related communities: a community that reinforces the disordered behavior and a community that helps them recover. The authors also discuss the interaction patterns between the two communities and the characteristics of the individual's behavior. Finally, Shuai et al. [30] extract a set of behavioral and social network structural features from users' online social network data and propose a new multi-source learning framework to identify individuals who are addicting to online social networks effectively.

The works mentioned above discuss and identify different social network structural factors that influence individuals' mental well-being, and models have been proposed to detect the potential mental disorder patients based on different social network features. Our work, which employs the network intervention strategy to modify the individuals' social network structure to improve their well-being, can be viewed as the next step toward helping those individuals with mental disorders.

Employing Computer Science techniques to help mental disorder patients has just started but is gaining more research attention. O'Leary et al. [31] conduct a two-week experiment with 40 people to discuss the effectiveness of talk therapy for mental disorder patients, under two different talk therapy settings: guided chats and unguided chats. They conclude that guided chats usually offer solutions to problems, while unguided chats can help experience sharing, and these insights can help design peer support chat systems. Moreover, Murnane et al. [32] study the social relations and interactions of 22 individuals with bipolar disorder and discuss the design implications for personal informatics systems. On the other hand, to help those with social network addiction, Shuai et al. [33] propose a new algorithm, based on behavioral therapy, to substitute addictive newsfeeds in a user's Facebook with some less addictive and more supportive newsfeeds, which is able to significantly reduce the users' addictive scores. Finally, Wilder et al. [34] discuss the relationship between social influence and obesity. They formulate a problem to optimize the social influence among the users. Although the above works aim at helping the people with mental disorder by employing techniques in Computer Science, they are different from the proposed research problem in this paper, and their approaches cannot be applied to our scenario directly. This is because they do not consider the important factor of small dense subgroups, and their algorithms are not designed to improve the well-being of the individuals under the concept of small dense subgroups.

4 Algorithm for SSMP with $\Delta = 3$

Although the general SSMP is NP-Hard, after carefully examining the problem, we observe that a special case of SSMP is indeed polynomial-time solvable, i.e., a polynomial-time algorithm exists that can obtain the optimal solution for SSMP with $\Delta = 3$. In the following, we first propose a polynomial-time algorithm for the special case of SSMP with $\Delta = 3$ in Section 4.1. Then, we improve the algorithm to linear-time with a specially designed data structure, *Ratio-Shifting Array (RSA)*, in Appendix C, available in the online supplemental material.

Please note that SSMP with $\Delta \leq 2$ is trivial. For $\Delta = 2$, the graph only contains a set of independent Δ-SGs (which are triangles), and these Δ-SGs are not overlapping or connected. In this case, the optimal solution is to randomly pick b vertices each from a unique Δ-SG and to link the b vertices to any vertex in A (the set of new vertices)[3]. On the other hand, for $\Delta = 1$ or $\Delta = 0$, no Δ-SG exists because for any vertex v in a triangle, $deg(v) \geq 2$ must hold, and Δ-SGs are inherently triangles.

For the ease of presentation, we assume that the input graph is preprocessed offline according to the given Δ to remove redundant vertices that will never contribute to the number of Δ-SGs. That is, the input graph G is preprocessed to construct $\widehat{G}$ by removing all the vertices which are i) not covered by any triangle or ii) with degrees greater than Δ. The preprocessing strategy does not change the optimal solution due to the following two reasons: i) If a vertex is not covered by any triangle, linking it to a vertex in A does not reduce Δ-SGs. ii) If a vertex v has degree greater than Δ, any triangle covering v is not a Δ-SG by definition. Therefore, linking v to any vertex in A does not reduce Δ-SGs as well.

4.1 Algorithm Design

We denote $deg_G(v)$ the original degree of vertex v in G, and denote $deg_{\widehat{G}}(v)$ the degree of v in the preprocessed graph $\widehat{G}$. Let c_v denote the *cost* of $v \in \widehat{G}$ where $c_v = \Delta - deg_G(v) + 1$. By definition, if the degree of a vertex covered by a Δ-SG exceeds Δ, the Δ-SG no longer exists. Therefore, c_v can be viewed as the number of edge additions required for v to eliminate the Δ-SGs covering v. Given a subset of vertices $S \subseteq \widehat{G}$, we denote $n^{\Delta}(v \ominus S)$ the number of Δ-SGs covering

3. If there are fewer than b Δ-SGs in G, say d Δ-SGs, then picking d vertices each from a unique Δ-SG and linking them to a vertex in A is the optimal solution.

TABLE 1
Summary of Notations Used by 3-SMMTG

Term	Description	Term	Description
c_v	$c_v = \Delta - deg_G(v) + 1$	$\widehat{G}$	Graph after preprocessing
$\mathbb{V}$	Vertex set after Cost Pruning	$n^{\Delta}(v \ominus S)$	Num. of Δ-SGs covering v but not S
$n^{\Delta}(S)$	Num. of Δ-SGs covering S	$\rho(V, E, F)$	$\delta(V, E) - \delta(V, E \cup F)$
A	New vertex set	F	New edge set

the vertex $v \in \widehat{G}$ but not covering any vertex in S. Moreover, $n^{\Delta}(S)$ is the number of Δ-SGs covering all the vertices in S. Table 1 summarizes the notations used in this section.

The proposed algorithm, namely *3-SMMTG*, works as follows. Given the preprocessed graph $\widehat{G}$, we first employ an effective pruning strategy, namely *Cost Pruning*, which partitions $\widehat{V}$ into two parts: $\mathbb{V}$ and the rest $\widehat{V} - \mathbb{V}$. Each vertex v in $\mathbb{V}$ satisfies both i) $c_v \leq b$ and ii) $c_v \leq s$. The pseudo-code of 3-SMMTG is listed in Algorithm 1, where line 1 of Algorithm 1 generates $\widehat{G}$ from G, and line 2 performs Cost Pruning on $\widehat{V}$ to construct $\mathbb{V}$.

In the subsequent steps, this algorithm only needs to examine the vertices in $\mathbb{V}$ and can safely skip the vertices in $\widehat{V} - \mathbb{V}$ to effectively reduce the computation time. The proof of the Cost Pruning strategy will be detailed later.

The core of the algorithm is to construct the *b-Max Set* $\mathbb{S}$ where $\mathbb{S} \subseteq \mathbb{V}$, such that the total cost of the vertices in $\mathbb{S}$ is smaller than b, i.e., $\sum_{u \in \mathbb{S}} c_u < b$, and the vertices in $\mathbb{S}$ are covered by the maximum number of Δ-SGs. The definition of b-Max Set is as follows.

Definition 2 (b-Max Set). *Given the parameters b and Δ, the b-Max Set $\mathbb{S}$ is the set of vertices in $\mathbb{V}$ such that $\sum_{u \in \mathbb{S}} c_u \leq b$, and $\mathbb{S}$ is covered by the maximum number of Δ-SGs in G.*

To construct the b-Max Set, 3-SMMTG first creates a set S and iteratively includes new vertices into S (line 3 of Algorithm 1). Later we will prove that the constructed S is indeed the b-Max Set. At the beginning, S is an empty set. Then, S is expanded by iteratively selecting vertices from $\mathbb{V}$ into S. To guide the algorithm for an effective exploration, 3-SMMTG adopts the notion of *cost ratio* for each vertex in respect to S. Given a vertex $v \in \mathbb{V}$, the cost ratio of v in respect to S is defined as $\frac{n^{\Delta}(v \ominus S)}{c_v}$, where $n^{\Delta}(v \ominus S)$ is the number of Δ-SGs covering v but not covering any vertex in S, and c_v is the cost of v. At each iteration (line 4 of Algorithm 1), 3-SMMTG extracts the vertex v with the maximum cost ratio among the vertices in $\mathbb{V}$ (line 5 of Algorithm 1). 3-SMMTG moves v into S if $\sum_{u \in S} c_u + c_v \leq b$ holds, otherwise v is truncated and 3-SMMTG enters another iteration (lines 6 to 8 of Algorithm 1). The iterations repeat until the total cost of the vertices in S is at least b, i.e., $\sum_{v \in S} c_v \geq b$.

After finding S, we construct A as a set with $\max_{v \in S} c_v$ vertices. Please note that $|A| = \max_{v \in S} c_v \leq s$ because $S \subseteq \mathbb{V}$ and any vertex $u \in \mathbb{V}$ satisfies $c_u \leq s$ after Cost Pruning. After A is constructed, for each vertex v in S, the algorithm links v to exactly c_v vertices in A, i.e., a_1, a_2, ..., a_{c_v}, and creates exactly c_v edges $(v, a_1), (v, a_2), ..., (v, a_{c_v})$ in F for each v. Finally, the algorithm returns the graph $G' = (V \cup A, E \cup F)$. These steps are shown in lines 11 to 13 of Algorithm 1.

Fig. 3 presents a step-by-step example for 3-SMMTG. Here, we have $b = 4$, $\Delta = 3$, and $s = 1$. Specifically, 3-SMMTG first preprocesses G, performs Cost Pruning (as stated in Lemma 1), and initializes the sets (lines 1 to 3 of Algorithm 1). The proposed Cost Pruning removes $v_5, v_6, v_7, v_9, v_{12}$, and we have $\mathbb{V} = \{v_1, v_2, v_3, v_4, v_8, v_{10}, v_{11}, v_{13}, v_{14}, v_{15}\}$.

3-SMMTG then finds the set S from Fig. 3a as follows. At the beginning, S is empty and the $U = \mathbb{V}$. In the first iteration, 3-SMMTG picks vertex v_1 into S, because v_1 incurs the maximum cost ratio (line 5 of Algorithm 1), i.e., $\frac{n^{\Delta}(v_1 \ominus S)}{c_{v_1}} = \frac{3}{1} = 3$. Next, v_1 is moved into S (lines 6 and 7 of Algorithm 1). Now, $S = \{v_1\}$, and $\sum_{u \in S} c_u = 1$. Then, v_1 is removed from U (line 8 of Algorithm 1). The vertex in S is shown as the dark vertex in Fig. 3b.

In the second iteration, the algorithm selects v_8, with $\frac{n^{\Delta}(v_8 \ominus S)}{c_{v_8}} = 2$ (line 5 of Algorithm 1). Please note that v_8, v_{11}, v_{13}, v_{15} all have the same cost ratio, i.e., 2, and the algorithms breaks the tie arbitrarily by selecting v_8. Afterward, v_8 is moved into S (lines 6 and 7 of Algorithm 1). Now, $S = \{v_1, v_8\}$, and $\sum_{u \in S} c_u = 2 < b = 4$. Next, v_8 is removed from U (line 8 of Algorithm 1). The dark vertices in Fig. 3c show the vertices in S after the second iteration.

In the third iteration, the algorithm selects v_{13}, with $\frac{n^{\Delta}(v_{13} \ominus S)}{c_{v_{13}}} = 2$, and removes v_{13} from U. Now, $S = \{v_1, v_8, v_{13}\}$, and $\sum_{u \in S} c_u = 3 < b = 4$. The dark vertices in Fig. 3d show the vertices in S after this iteration.

Finally, 3-SMMTG chooses v_2 (line 5 of the Algorithm 1) with $\frac{n^{\Delta}(v_2 \ominus S)}{c_{v_2}} = \frac{1}{1} = 1$. After this step, $S = \{v_1, v_8, v_{13}, v_2\}$, and

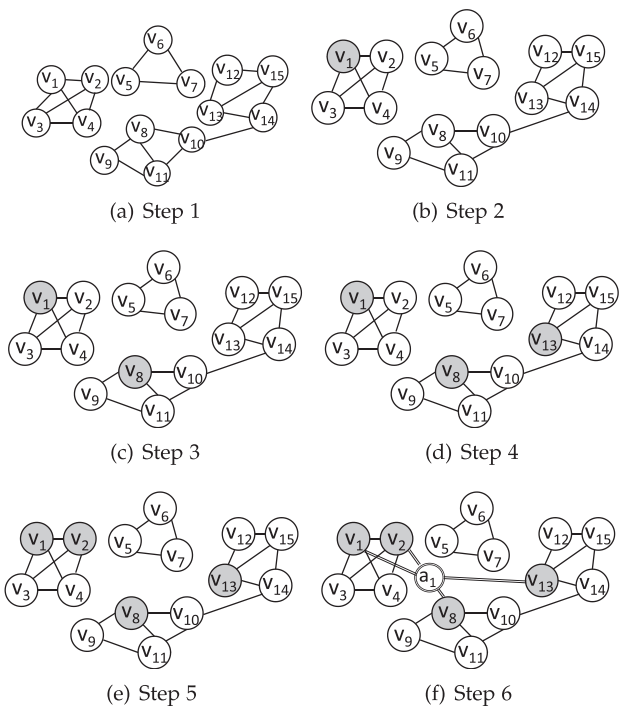


Fig. 3. Step-by-step example of 3-SMMTG.

$\sum_{u\in S} c_u = 4 \geq b$. Therefore, 3-SMMTG stops including new vertices into S (line 4 of Algorithm 1). The dark vertices in Fig. 3e are the vertices in S after this iteration.

After finding the set S, the algorithm constructs the set A accordingly. As $\max_{v\in S} c_v = 1$, 3-SMMTG creates one vertex in A, namely a_1 (line 11 of Algorithm 1). In the next step, the algorithm creates new edges linking from the vertices in S to A (line 12 of Algorithm 1). Specifically, for vertices v_1, v_2, v_8, and v_{13}, their costs are all 1. Therefore, one edge is created by linking each of $\{v_1, v_2, v_8, v_{13}\}$ to a_1. Finally, the algorithm outputs the intervened graph (line 13 of Algorithm 1). The resulting intervened graph is shown in Fig. 3f and it reduces 8 Δ-SGs, which is the optimal solution to this instance.

Now, we consider the pruning strategy mentioned above, i.e., Cost Pruning, which is able to significantly reduce the number of vertices that need to be considered. Specifically, the Cost Pruning strategy is elaborated in Lemma 1.

Lemma 1 (Cost Pruning). *For any vertex $v \in \widehat{G}$, v can be pruned if $c_v > b$ or $c_v > s$.*

Proof. If $c_v > b$, we do not have to consider linking v to any vertex in A because even if we link all the b new edges to v (i.e., linking b edges from v to b vertices in A), the degree of v will still be smaller than $\Delta + 1$, indicating that no Δ-SGs disappears.

Similarly, if $c_v > s$, there will not be enough new vertices to allow vertex v to have degree greater than Δ, implying that increasing the degree of v cannot reduce any Δ-SG. Therefore, such vertices can be skipped (but cannot be removed from G because the edges linking from them are still meaningful). □

Algorithm 1. 3-SMMTG

Input: Graph $G = (V, E)$ with parameters b, s ($\Delta = 3$)
Output: Graph $G' = (V \cup A, E \cup F)$ with maximized $\rho(V, E, F)$
1: Process G to $\widehat{G} = (\widehat{V}, \widehat{E})$, which contains only Δ-SGs
2: Perform Cost Pruning on $\widehat{V}$ to obtain $\mathbb{V}$
3: $S \leftarrow \varnothing$; $A \leftarrow \varnothing$; $F \leftarrow \varnothing$; $U \leftarrow \mathbb{V}$;
4: **while** $\sum_{u\in S} c_u < b$ **do**
5: Select $v \in U$ that maximizes $\frac{n^{\Delta}(v \ominus S)}{c_v}$ (ties can be broken arbitrarily)
6: **if** $\sum_{u\in S} c_u + c_v \leq b$ **then**
7: $S \leftarrow S \cup \{v\}$
8: $U \leftarrow U \setminus \{v\}$
9: **if** $U = \varnothing$ **then**
10: Break;
11: Construct A with $\max_{u\in S} c_u$ vertices
12: For each vertex v in S, link v to c_v vertices $\{a_1, a_2, \ldots, a_{c_v}\} \subseteq A$ and create the corresponding edges $(v, a_1), \ldots, (v, a_{c_v})$ in F
13: **return** $G' = (V \cup A, E \cup F)$

4.2 Analysis of 3-SMMTG

In the following, we prove that 3-SMMTG obtains the optimal solution for SSMP with $\Delta = 3$ and analyze its time complexity. We first prove that given a b-Max Set as defined in Definition 2, we can always maximize $\rho(V, E, F)$. Then, we show that there are only 4 different cost ratios when $\Delta = 3$ and prove that the set S obtained by 3-SMMTG satisfies Definition 2, i.e., S is a b-Max Set. Finally, Theorem 3 proves that 3-SMMTG obtains the optimal solution in polynomial time.

Specifically, given a b-Max Set $\mathbb{S}$, the following lemma shows that we can obtain the minimum $\delta(V, E \cup F)$ to maximize $\rho(V, E, F)$.

Lemma 2. *For an instance of SSMP with parameters b and s, if b-Max Set $\mathbb{S} \subseteq \mathbb{V}$ is obtained, the optimal solution of the SSMP instance can be obtained.*

Proof. Given a b-Max Set $\mathbb{S} \subseteq \mathbb{V}$, we can find $G' = (V \cup A, E \cup F)$ such that $deg_{G'}(v) > \Delta, \forall v \in S$, where the new vertex set A and the new edge set F are constructed as follows. The vertex set A includes exactly $\max_{u\in\mathbb{S}} c_u$ vertices. For each v in $\mathbb{S}$, we link v to each vertex in $\{a_1, a_2, \ldots, a_{c_v}\} \subseteq A$ and create the edges $\{(v, a_1), (v, a_2), \ldots, (v, a_{c_v})\}$ in F. This procedure ensures that the degree of each vertex in $\mathbb{S}$ is greater than Δ.

Since the set $\mathbb{S}$ is covered by the largest number of Δ-SGs in G ($\mathbb{S}$ is the set of vertices that maximize $n^{\Delta}(\mathbb{S})$), making the degree of all the vertices in $\mathbb{S}$ be greater than Δ reduces the largest number of Δ-SGs, i.e., minimizes $\delta(V, E \cup F)$. This is because $\delta(V, E \cup F) = \delta(V, E) - n^{\Delta}(\mathbb{S})$. We prove this with contradiction. If there exists another set $T \neq \mathbb{S}$ such that making the degrees of all the vertices in T be greater than Δ results in a smaller $\delta(V, E \cup F)$, then $n^{\Delta}(\mathbb{S}) < n^{\Delta}(T)$ must hold. Since $\mathbb{S}$ maximizes $n^{\Delta}(\mathbb{S})$ and $\mathbb{S}$ satisfies $\sum_{v\in\mathbb{S}} c_v \leq b$, $\sum_{v\in T} c_v > b$ holds. Therefore, the budget b is not sufficiently large for intervening all vertices in T to have degrees greater than Δ. This leads to a contradiction. Therefore, such T does not exist, and making the degrees of all the vertices in S be greater than Δ minimizes $\delta(V, E \cup F)$, which implies that $\rho(V, E, F)$ is maximized. The lemma follows. □

From the direct result of Lemma 2, we only describe how to find b-Max Set $\mathbb{S}$ in the following, because the optimal solution of SSMP can be obtained easily if the b-Max Set $\mathbb{S}$ is found. Now we prove that the proposed 3-SMMTG can indeed obtain the b-Max Set $\mathbb{S}$ in polynomial time. We first analyze the cost ratios in Lemma 3.

Lemma 3. *For any vertex $v \in \mathbb{V}$, the cost ratio of v must be either 0.5, 1, 2, or 3. That is, $\frac{n^{\Delta}(v \ominus S)}{c_v} \in \{0.5, 1, 2, 3\}$ holds.*

Proof. Since $\Delta = 3$, the maximum number of triangles covering any vertex is 3. When a vertex v is covered by 3 or 2 triangles, v's degree can only be 3, indicating that $c_v = \Delta - deg_G(v) + 1 = 1$ holds. At the same time, the cost ratio of v is 3 (covered by 3 triangles) or 2 (covered by 2 triangles). For another case, when a vertex v is covered by 1 triangle, the degree of v can be 2 ($c_v = 2$) or 3 ($c_v = 1$). Therefore, the cost ratio is either 0.5 or 1 in this case. □

Now, we prove that the set S obtained by 3-SMMTG is indeed a b-Max Set.

Lemma 4. *The set S obtained by 3-SMMTG is a b-Max Set.*

Proof. Lemma 3 shows that there are only 4 different cost ratios. Here, we categorize them into 4 different types of triangle structures, as shown in Fig. 4. Figs. 4a and 4b illustrate the vertex (black) with cost ratio 3 (type-1) and 2 (type-2), respectively. Fig. 4c shows the vertex with cost ratio 1 (type-3) and Fig. 4d shows the vertex with cost ratio 0.5 (type-4).

Since $\Delta = 3$, the maximum degree of each vertex is 3, indicating that the 4 structures of Fig. 4 do not share any vertex with other types. If all vertices in S are not included in a structure, we say that structure is an unselected structure (all vertices in that structure are not selected into S).

Here, we consider the different scenarios when selecting different numbers of vertices in a structure into S. Recall that all structures are independent to each other, so we can consider each structure independently. If we select a vertex in a type-1 structure into S, this makes its neighbors' cost ratios be 1. Therefore, after a vertex in type-1 structure is selected into S, its neighbors can only be selected into S if there is no remaining vertices in type-2 structures. Moreover, if there are two type-1 vertices included in S, the remaining vertices must have cost ratio 0, and thus these vertices do not need to be considered in the future. If we select a vertex in a type-2, type-3, or type-4 structure into S, the cost ratios of the vertex's neighbors becomes 0. Therefore, these neighbors do not need to be considered in the future.

We prove this lemma with an induction on b. When $b = 1$, 3-SMMTG picks the vertex with the maximum cost ratio into the empty S. This operation maximizes $n^{\Delta}(S)$. We then assume that when $b = k - 1$, the statement of the lemma holds.

When $b = k$, we need to prove that the algorithm is able to maximize $n^{\Delta}(S)$ where $\sum_{u \in S} c_u \leq k$. In the algorithm, we assume that we can find S that maximizes $n^{\Delta}(S)$ with $\sum_{u \in S} c_u \leq k - 1$. As the triangle structures are independent, the vertices we select into S do not affect the cost ratios of the other vertices in other triangle structures. Since only the vertices with cost ratio 0.5 can have the cost 2, if vertex v's cost ratio is maximum, then its cost c_v must be minimum. This is because when $\Delta = 3$, only the vertices with cost ratio 0.5 can have cost equal to 2, which is the maximum cost (the cost can only be 1 or 2 in this case), and 0.5 is the minimum cost ratio when $\Delta = 3$. Therefore, all the vertices which have been selected into S have smaller or equal cost to the vertices that are not yet selected into S.

In the following, let v_{max} denote the vertex with the maximum cost ratio in $\mathbb{V} \setminus S$. If v_{max} is in the structures that do not include any vertices in S, and $c_{v_{max}} < k - \sum_{u \in S} c_u$, selecting v_{max} into S maximizes $n^{\Delta}(S)$ because the cost ratio of v_{max} is not affected by the vertices in S.

If v_{max} is in the structures that include some vertices in S, and $c_{v_{max}} < k - \sum_{u \in S} c_u$, the maximum cost ratio will be 1 and the cost $c_{v_{max}}$ is equal to 1. From the observation above, we know that all the type-2 structures must cover the vertices in S. Thus selecting v_{max} into S maximizes $n^{\Delta}(S)$.

If v_{max} has cost $c_{v_{max}} > k - \sum_{u \in S} c_u$, S already incurs the maximum $n^{\Delta}(S)$ because there is no other vertex in $\mathbb{V} \setminus S$ that can be selected into S according to the cost ratio property mentioned above, where the property states that if the vertex v's cost ratio is maximum, its cost c_v must be minimum. This implies that $c_{v_{max}}$ is the minimum cost among the vertices' costs in $\mathbb{V} \setminus S$. Therefore, no other vertex in $\mathbb{V} \setminus S$ can be selected into S. In summary, when $b = k$, the statement of the lemma holds. The lemma follows. □

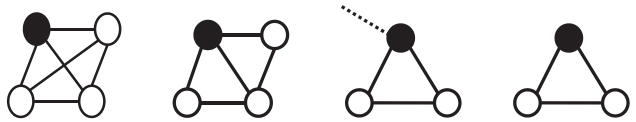


Fig. 4. Four different cost ratios.

The following theorem summarizes the above theoretical results and analyzes the time complexity of the proposed 3-SMMTG algorithm. We further improve the time complexity of the proposed 3-SMMTG to *linear time*, i.e., $O(|V|)$ time, with a specially designed data structure, *Ratio-Shifting Array*, in Appendix C, available in the online supplemental material.

Theorem 3. *3-SMMTG obtains the optimal solution to SSMP with $\Delta = 3$ in $O(|V|^2)$ time.*

Proof. Lemma 2 states that if the b-Max Set $\mathbb{S}$ is obtained, the optimal solution can be constructed. From Lemma 3, we know that 3-SMMTG can obtain b-Max Set, i.e., S. Therefore, 3-SMMTG obtains the optimal solution to SSMP with Δ=3.

We now analyze the time complexity. We consider the input graph $G = (V, E)$ already preprocessed offline such that G contains only feasible Δ-SGs. First, the algorithm spends $O(|V|)$ time to perform Cost Pruning. Since we can select at most $|V|$ vertices, the while loop from lines 4 to 10 in Algorithm 1 performs at most $|V|$ iterations. Inside the while loop, it needs $deg_G(v) \leq 3 = O(1)$ time to update the cost ratios and $O(|V|)$ time to find the vertex with the maximum cost ratio. Therefore, the time complexity for the while loop is $O(|V|^2)$. For the construction of G' (more precisely, A and F) in lines 11 to 12, the algorithm constructs at most 2 new vertices in A and adds at most b new edges. Since the cost of each vertex in $\mathbb{V}$ is at most 2 when $\Delta = 3$, b is at most $2 \cdot |V|$. The running time for constructing G' is $O(|V|)$. The overall time complexity of 3-SSMTG is $O(|V|^2)$. □

The bottleneck of the naïve implementation of 3-SMMTG is the extraction of the vertex with the maximum cost ratio, which takes $O(|V|)$ time for each extraction. However, as there are only 4 different cost ratios as stated in Lemma 3, we propose a specially designed data structure, named *Ratio-Shifting Array (RSA)* to reduce the time of each extraction from $O(|V|)$ to $O(1)$. With RSA, we are able to improve the performance of 3-SMMTG and reduce the time complexity to $O(|V|)$. Please refer to Appendix C, available in the online supplemental material, for the details of the proposed Ratio-Shifting Array.

5 Approximation Algorithm for General SSMP

In this section, we propose an approximation algorithm for the general SSMP, namely *ESGR*, which is an efficient algorithm that achieves an approximation ratio of $\frac{1}{2} \cdot (1 - \frac{1}{e})$. We detail the algorithm design of ESGR and formally prove its

TABLE 2
Summary of Notations Used by ESGR

Term	Description	Term	Description
v_i	The ith vertex selected into S	$\iota(v_i)$	The iteration that the ith vertex is selected into S
U_k	$\mathbb{V} \setminus \{v_1, v_2, \ldots, v_k\}$	$n^{\Delta}(\mathbb{S} \ominus S_{i-1})$	Num. of Δ-SGs covering $\mathbb{S}$ but not S_{i-1}
$v_{\mathbb{S}}^1$	The first vertex in $\mathbb{S}$ selected by ESGR in iteration $\iota(v_{\mathbb{S}}^1)$ but not added to S	$\vartheta(v_{\mathbb{S}}^1)$	Num. of vertices selected to S before $v_{\mathbb{S}}^1$

approximation ratio. To boost the performance of ESGR, we also propose effective pruning strategies to avoid unnecessary examinations of vertices. Table 2 summarizes the notations used in this section.

5.1 Algorithm Description of ESGR

The basic idea of ESGR is similar to the greedy approach in 3-SMMTG. However, simply employing such a greedy approach cannot achieve a guaranteed performance bound for the general SSMP. We illustrate with an example below. In Fig. 5a, given $\Delta = 10$, $b = 4$, and $s = 4$. Vertices v_1 and v_2 are considered to be selected into S. The cost of v_1 is b (i.e., 4), and v_1 is covered by $(b-2)$ Δ-SGs; v_2 has cost 1 and is covered by exactly 1 Δ-SG. The 3-SMMTG algorithm picks v_2 into S first because v_2 has a larger cost ratio, i.e., $\frac{1}{1}$. In the next iteration, 3-SMMTG finds that the budget runs out and it cannot select v_1 into S (the total cost of v_1 and v_2 is 5). Therefore, 3-SMMTG stops selecting any other vertex into S. However, the optimal solution is to select v_1 into S, which is covered by 2 Δ-SGs. In this case, the approximation ratio is $\frac{1}{b-2}$, which is unbounded because b can be set as any arbitrary positive integer.

In fact, the above example manifests that if there exists a vertex with a high cost and is covered by many Δ-SGs, then this vertex will have a low priority when the greedy approach selects vertices according to their cost ratios. However, if such a vertex is selected, it is indeed covered by more Δ-SGs. To address the problem mentioned above, we devise algorithm ESGR by considering the existence of such vertices and prove that ESGR is a constant-ratio approximation algorithm to SSMP, i.e., the approximation ratio is $\frac{1}{2} \cdot (1 - \frac{1}{e})$.

Specifically, ESGR first finds the set S by iteratively selecting the vertex with the maximum cost ratio into S, similar to 3-SMMTG. That is, the set S is empty initially (line 3 of Algorithm 2). In each iteration, to find the suitable vertex for S, ESGR identifies the vertex $v' \notin S$ with the maximum cost ratio and moves it into S if $\sum_{u \in S} c_u + c_{v'} \leq b$ holds (lines 5 to 7 of Algorithm 2). The procedure repeats while the sum of the vertex costs in S is less than b, i.e., $\sum_{u \in S} c_u < b$. ESGR regards S as the first candidate set. In addition, to effectively consider those vertices that come with high costs and are covered by many Δ-SGs at the same time, ESGR generates the second candidate set as follows (line 11 of Algorithm 2). The second candidate set is the set containing only one single vertex, i.e., $\{v_{max}\}$, where $v_{max} = \arg\max_{v \in V} n^{\Delta}(v)$. ESGR then extracts the candidate set that is covered by more Δ-SGs from the two candidate sets to construct the output graph G' (lines 12 to 16 of Algorithm 2). As will be proved, the identification of the second candidate set is a crucial step to achieve the guaranteed performance bound. The pseudocode of ESGR is presented in Algorithm 2.

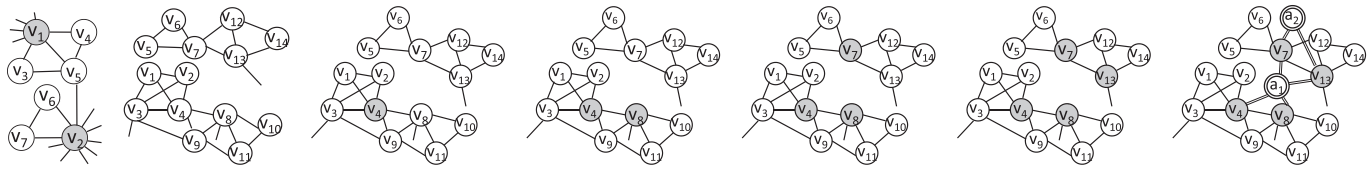
Algorithm 2. ESGR

Input: Graph $G = (V, E)$ with parameters b, s, Δ
Output: Graph $G' = (V \cup A, E \cup F)$ with maximized $\rho(V, E, F)$
1: Process G to $\widehat{G} = (\widehat{V}, \widehat{E})$, which contains only Δ-SGs
2: Perform Cost Pruning on $\widehat{V}$ to obtain $\mathbb{V}$
3: $S \leftarrow \varnothing$; $A \leftarrow \varnothing$; $F \leftarrow \varnothing$; $U \leftarrow \mathbb{V}$;
4: **while** $\sum_{u \in S} c_u < b$ **do**
5: Find $v' \in U$ that maximizes $\frac{n^{\Delta}(v' \ominus S)}{c_{v'}}$ (ties can be broken arbitrarily)
6: **if** $\sum_{u \in S} c_u + c_{v'} \leq b$ **then**
7: $S \leftarrow S \cup \{v'\}$
8: $U \leftarrow U \setminus \{v'\}$
9: **if** $U = \varnothing$ **then**
10: Break;
11: Select $v_{max} \in \mathbb{V}$ that maximizes $n^{\Delta}(\{v_{max}\})$
12: **if** $n^{\Delta}(\{v_{max}\}) > n^{\Delta}(S)$ **then**
13: $S \leftarrow \{v_{max}\}$
14: Construct A with $\max_{u \in S} c_u$ vertices
15: For each vertex v in S, link v to c_v vertices $\{a_1, a_2, \ldots, a_{c_v}\} \subseteq A$ and create the edges $(v, a_1), (v, a_2), ...(v, a_{c_v})$ in F
16: **return** $G' = (V \cup A, E \cup F)$

Figs. 5b, 5c, 5d, 5e, 5f, and 5g present a step-by-step example of ESGR with $\Delta = 5$, $s = 2$, and $b = 6$. ESGR first preprocesses G into $\widehat{G}$, performs the Cost Pruning, and initializes the sets (lines 1 to 3 of Algorithm 2). After Cost Pruning, $v_1, v_2, v_5, v_6, v_{10}, v_{11}, v_{12}, v_{14}$ are pruned off, and we have $\mathbb{V} = \{v_3, v_4, v_7, v_8, v_9, v_{13}\}$.

ESGR starts to construct the first candidate set (lines 4 to 10). ESGR extracts the vertex with the maximum cost ratio in each iteration into S. ESGR first moves v_4 into S since v_4 has the maximum cost ratio (i.e., 5) (line 5 of Algorithm 2). Next, U is updated (line 8 of Algorithm 2). After this iteration, the vertex selected in S is shown as the dark vertex in Fig. 5c. In the second iteration, ESGR selects v_8 since v_8 has the maximum cost ratio (i.e., 2) (line 5 of Algorithm 2). Now, $S = \{v_4, v_8\}$, and the total cost of S is $\sum_{v \in S} c_v = 2 \leq b$. The dark vertices in Fig. 5d show the vertices in S after this iteration.

In the subsequent two iterations, ESGR selects v_7, and v_{13} because they incur the maximum cost ratios, (line 5 of the Algorithm 2), i.e., 1. Here, since v_7 and v_{13} all have the same cost ratio, the tie is broken by ESGR arbitrarily. Then, the algorithm updates U by removing v_7 and v_{13} from it (line 8 of Algorithm 2). The vertices in S after these two iterations are shown in Figs. 5e and 5f, respectively. Now, $S = \{v_4, v_8, v_7, v_{13}\}$, and the total cost of S is $\sum_{v \in S} c_v = 6 \geq b$. ESGR stops moving vertices into S (line 4 of Algorithm 2). For the second candidate set, ESGR finds that vertex v_4 is covered by the



(a) Counter-example (b) Step 1 (c) Step 2 (d) Step 3 (e) Step 4 (f) Step 5 (g) Step 6

Fig. 5. Step-by-step example of ESGR.

maximum number of Δ-SGs. Therefore, ESGR sets $\{v_4\}$ as the second candidate set (line 11 of Algorithm 2).

Finally, ESGR compares the two candidate sets and extracts the one that is covered by more Δ-SGs. Since the first candidate set S is covered by 10 Δ-SGs (whereas the second candidate set $\{v_4\}$ is covered by 5 Δ-SGs), ESGR selects $S = \{v_4, v_8, v_7, v_{13}\}$ to construct the output graph G' (lines 12 to 13 of Algorithm 2). The construction of G' is the same as that in 3-SMMTG (lines 14 to 16 of Algorithm 2). Fig. 5g presents the resulting G'.

5.2 Analysis of Approximation Ratio for ESGR

In the following, we analyze the approximation ratio of ESGR. Please note that Lemma 2 in Section 4 indicates that if we find a b-Max Set $\mathbb{S}$, we are able to obtain the optimal solution for SSMP. That is, the number of Δ-SGs that can be reduced in the optimal solution is equal to $n^{\Delta}(\mathbb{S})$, i.e., the number of Δ-SGs covering the b-Max Set $\mathbb{S}$. In other words, based on Lemma 2, the ratio of $n^{\Delta}(S)$ to $n^{\Delta}(\mathbb{S})$ equals the approximation ratio. Therefore, in our proofs below, we focus on deriving the guaranteed ratio of $\frac{n^{\Delta}(S)}{n^{\Delta}(\mathbb{S})}$.

Given a b-Max Set $\mathbb{S}$, at each iteration, let v' be the vertex to be added to S (i.e., the vertex that satisfies $\sum_{u \in S} c_u + c_{v_i} \leq b$ in line 6 of Algorithm 2). Moreover, we denote $\iota(v')$ the iteration of the while loop in line 4 of Algorithm 2 such that v' is added into S.

Let $v^1_{\mathbb{S}}$ be the first vertex that belongs to $\mathbb{S}$ which is selected by ESGR in iteration $\iota(v^1_{\mathbb{S}})$, but is not added to S due to the budget constraint. That is, $\iota(v^1_{\mathbb{S}})$ is the first iteration such that: i) $v^1_{\mathbb{S}} \in \mathbb{S}$, ii) $v^1_{\mathbb{S}}$ incurs the maximum cost ratio in iteration $\iota(v^1_{\mathbb{S}})$, and iii) $\sum_{u \in S} c_u + c_{v^1_{\mathbb{S}}} > b$ at iteration $\iota(v^1_{\mathbb{S}})$. Let $\vartheta(v^1_{\mathbb{S}})$ denote the number of vertices that have been selected into S before $v^1_{\mathbb{S}}$ is considered. Please note that if the vertex $v^1_{\mathbb{S}}$ does not exist, it implies that $\mathbb{S} \subseteq S$. If $\mathbb{S} = S$, then ESGR obtains the optimal solution. If $\mathbb{S} \subset S$, it indicates that ESGR obtains a solution better than the optimal solution, which is a contradiction. Therefore, in our analysis, we can assume that $v^1_{\mathbb{S}}$ always exists. Moreover, to better analyze the algorithm, let $i = \iota(v')$, we also denote S_i the set of vertices selected through iteration 1 to iteration $\iota(v')$, and let $U_i = \mathbb{V} \setminus S_i$. In other words, $S_i \cup U_i = \mathbb{V}$ always holds.

In the following, we first derive the relationship between the numbers of Δ-SGs covering S_i and S_{i-1}.

Lemma 5. *For any iteration before iteration* $\iota(v^1_{\mathbb{S}})$, $n^{\Delta}(S_i) - n^{\Delta}(S_{i-1}) \geq \frac{c_{v'}}{b} \cdot n^{\Delta}(\mathbb{S}) - \frac{c_{v'}}{b} \cdot n^{\Delta}(S_{i-1})$ *must hold for* $1 \leq i \leq \vartheta(v^1_{\mathbb{S}}) + 1$.

Proof. Let $n^{\Delta}(\mathbb{S} \ominus S_{i-1})$ be the number of Δ-SGs which cover $\mathbb{S}$ but do not cover S_{i-1}. We observe that $n^{\Delta}(\mathbb{S} \ominus S_{i-1})$ is at least $n^{\Delta}(\mathbb{S}) - n^{\Delta}(S_{i-1})$. The reason is as follows. Let X denote the set of Δ-SGs covering $\mathbb{S}$, and let Y denote the set of Δ-SGs covering S_{i-1}. $|X| - |X \cap Y| \geq |X| - |Y|$ must hold. Please note that, if there exist vertices discarded from U before choosing v', then they are not in $\mathbb{S}$ (by definition of $v^1_{\mathbb{S}}$). Since v' is the vertex v with the maximum cost ratio among the vertices in U_{i-1} such that the total cost of $S_{i-1} \cup \{v\}$ is at most b, we have that, for each vertex in $\mathbb{S} \setminus S_{i-1}$, its cost ratio is smaller than or equal to the cost ratio of v', i.e., $\frac{n^{\Delta}(v' \ominus S_{i-1})}{c_{v'}}$.

Therefore, the cost ratios of the vertices in $\mathbb{S} \setminus S_{i-1}$ are at most $\frac{n^{\Delta}(v' \ominus S_{i-1})}{c_{v'}}$. Since the total cost of $\mathbb{S}$ is at most b, $\sum_{u \in \mathbb{S} \setminus S_{i-1}} c_u \leq \sum_{u \in \mathbb{S}} c_u \leq b$ holds, and $n^{\Delta}(\mathbb{S} \setminus S_{i-1})$ is at most $b \cdot \frac{n^{\Delta}(v' \ominus S_{i-1})}{c_{v'}}$, which is greater than or equal to $n^{\Delta}((\mathbb{S} \setminus S_{i-1}) \ominus S_{i-1})$.

Since $n^{\Delta}(\mathbb{S} \ominus S_{i-1}) = n^{\Delta}((\mathbb{S} \setminus S_{i-1}) \ominus S_{i-1})$, we have $b \cdot \frac{n^{\Delta}(v' \ominus S_{i-1})}{c_{v'}} \geq n^{\Delta}((\mathbb{S} \setminus S_{i-1}) \ominus S_{i-1})$ because the cost ratios of the vertices in $\mathbb{S} \setminus S_{i-1}$ are at most $\frac{n^{\Delta}(v' \ominus S_{i-1})}{c_{v'}}$, and the total cost of $\mathbb{S} \setminus S_{i-1}$ is at most b. Moreover, $n^{\Delta}((\mathbb{S} \setminus S_{i-1}) \ominus S_{i-1}) = n^{\Delta}(\mathbb{S} \ominus S_{i-1}) \geq n^{\Delta}(\mathbb{S}) - n^{\Delta}(S_{i-1})$ as described in the beginning of this proof.

Please note that $n^{\Delta}(v_i \ominus S_{i-1}) = n^{\Delta}(S_i) - n^{\Delta}(S_{i-1})$ holds. By substituting $n^{\Delta}(v_i \ominus S_{i-1})$ with $n^{\Delta}(S_i) - n^{\Delta}(S_{i-1})$, we have $b \cdot \frac{n^{\Delta}(S_i) - n^{\Delta}(S_{i-1})}{c_{v'}} \geq n^{\Delta}(\mathbb{S}) - n^{\Delta}(S_{i-1})$. Then, we multiply both sides with $\frac{c_{v'}}{b}$ and we have $n^{\Delta}(S_i) - n^{\Delta}(S_{i-1}) \geq \frac{c_{v'}}{b} \cdot n^{\Delta}(\mathbb{S}) - \frac{c_{v'}}{b} \cdot n^{\Delta}(S_{i-1})$. □

Then, we derive the relationship between $n^{\Delta}(S_i)$ and $n^{\Delta}(\mathbb{S})$ in the following lemma.

Lemma 6. *For any iteration before iteration* $\iota(v^1_{\mathbb{S}})$, $n^{\Delta}(S_i) \geq [1 - \prod_{k=1}^{i}(\frac{b - c_{v_k}}{b})] \cdot n^{\Delta}(\mathbb{S})$ *must hold for* $1 \leq i \leq \vartheta(v^1_{\mathbb{S}}) + 1$.

Proof. We prove this lemma by induction. After the first vertex v_1 is selected into S, i.e., $S_1 = \{v_1\}$, $n^{\Delta}(S_1) = n^{\Delta}(v_1 \ominus \varnothing)$, i.e., the number of Δ-SGs covering S_1 is equal to the number of Δ-SGs covering v_1. We need to prove that $n^{\Delta}(v_1 \ominus \varnothing) \geq \frac{c_{v_1}}{b} \cdot n^{\Delta}(\mathbb{S})$. This is true because the cost ratio $\frac{n^{\Delta}(v_1 \ominus \varnothing)}{c_{v_1}}$ for v_1 is maximum over all the vertices in $\mathbb{V}$ and the maximum cost is at most the budget b.

Assume that when the (i-1)th vertex v_{i-1} is considered, the lemma holds. Now, we show that the lemma holds when the ith vertex v_i is considered. First, we add $(n^{\Delta}(S_{i-1}) - n^{\Delta}(S_{i-1}))$ to $n^{\Delta}(S_i)$ to make it more obvious to apply Lemma 5. Therefore, $n^{\Delta}(S_i) = n^{\Delta}(S_i) + (n^{\Delta}(S_{i-1}) - n^{\Delta}(S_{i-1})) = (n^{\Delta}(S_i) - n^{\Delta}(S_{i-1})) + n^{\Delta}(S_{i-1})$. From the above equation and Lemma 5, we have $n^{\Delta}(S_i) \geq \frac{c_{v_i}}{b} \cdot n^{\Delta}(\mathbb{S}) - \frac{c_{v_i}}{b} \cdot n^{\Delta}(S_{i-1}) + n^{\Delta}(S_{i-1}) = \frac{c_{v_i}}{b} \cdot n^{\Delta}(\mathbb{S}) + \frac{b - c_{v_i}}{b} \cdot n^{\Delta}(S_{i-1})$.

Moreover, from the induction assumption, we have $n^{\Delta}(S_i) \geq \frac{c_{v_i}}{b} \cdot n^{\Delta}(\mathbb{S}) + \frac{b-c_{v_i}}{b} \cdot [1 - \prod_{k=1}^{i-1}(\frac{b-c_{v_k}}{b})] \cdot n^{\Delta}(\mathbb{S}) = n^{\Delta}(\mathbb{S}) - \frac{b-c_{v_i}}{b} \cdot [\prod_{k=1}^{i-1}(\frac{b-c_{v_k}}{b})] \cdot n^{\Delta}(\mathbb{S}) = [1 - \prod_{k=1}^{i}(\frac{b-c_{v_k}}{b})] \cdot n^{\Delta}(\mathbb{S})$. The lemma follows. □

Theorem 4. *Given a b-Max Set $\mathbb{S}$ and the S obtained by ESGR, $\frac{n^{\Delta}(S)}{n^{\Delta}(\mathbb{S})} = \frac{1}{2} \cdot (1 - \frac{1}{e})$, i.e., ESGR is a $\frac{1}{2} \cdot (1 - \frac{1}{e})$-approximation algorithm to SSMP.*

Proof. Assume that ESGR is at iteration $\iota(v_{\mathbb{S}}^1)$, from Lemma 6, we have $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)+1}) \geq [1 - \prod^{\vartheta(v_{\mathbb{S}}^1)+ \ 1_{k=1}}(\frac{b-c_{v_k}}{b})] \cdot n^{\Delta}(\mathbb{S}) = [1 - \prod_{k=1}^{\vartheta(v_{\mathbb{S}}^1)+1}(1 - \frac{c_{v_k}}{b})] \cdot n^{\Delta}(\mathbb{S})$. Since adding $v_{\mathbb{S}}^1$ to $S_{\vartheta(v_{\mathbb{S}}^1)}$ violates the constraint b, $\sum_{v \in S_{\vartheta(v_{\mathbb{S}}^1)}} c_v + c_{v_{\mathbb{S}}^1} > b$ holds. We have $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)+1}) \geq [1 - \prod_{k=1}^{\vartheta(v_{\mathbb{S}}^1)+1}(1 - \frac{c_{v_k}}{\sum_{v \in S_{\vartheta(v_{\mathbb{S}}^1)+1}} c_v})] \cdot n^{\Delta}(\mathbb{S})$.

Next, recall that for any positive real numbers $c_1, c_2, \ldots, c_n$ where $\sum_{i=1}^{n} c_i = C$, the function $(1 - \prod_{i=1}^{n}(1 - \frac{c_i}{C}))$ is minimum when $c_1 = c_2 = ... = c_n = \frac{C}{n}$. Therefore, we derive the following inequality: $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)+1}) \geq [1 - (1 - \frac{1}{\vartheta(v_{\mathbb{S}}^1)+1})^{\vartheta(v_{\mathbb{S}}^1)+1}] \cdot n^{\Delta}(\mathbb{S}) \geq (1 - \frac{1}{e}) \cdot n^{\Delta}(\mathbb{S})$. As $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)+1}) = n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)}) + n^{\Delta}(v_{\mathbb{S}}^1 \ominus S_{\vartheta(v_{\mathbb{S}}^1)})$ holds, from the above inequality, we have $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)}) + n^{\Delta}(v_{\mathbb{S}}^1 \ominus S_{\vartheta(v_{\mathbb{S}}^1)}) \geq (1 - \frac{1}{e}) \cdot n^{\Delta}(\mathbb{S})$.

Then, we observe that $n^{\Delta}(v_{\mathbb{S}}^1 \ominus S_{\vartheta(v_{\mathbb{S}}^1)})$ is at most the maximum number of the Δ-SGs covering a single vertex $v \in \mathbb{V}$. In other words, the number of Δ-SGs covering the second candidate set $\{v_{max}\}$, i.e., $n^{\Delta}(\{v_{max}\})$ in ESGR is at least $n^{\Delta}(v_{\mathbb{S}}^1 \ominus S_{\vartheta(v_{\mathbb{S}}^1)})$. Therefore, $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)}) + n^{\Delta}(\{v_{max}\}) \geq (1 - \frac{1}{e}) \cdot n^{\Delta}(\mathbb{S})$. From the inequality above, either $n^{\Delta}(S_{\vartheta(v_{\mathbb{S}}^1)})$ (i.e., the number of Δ-SGs covering the first candidate set) or $n^{\Delta}(\{v_{max}\})$ (i.e., the number of Δ-SGs covering the second candidate set) is at least $\frac{1}{2} \cdot (1 - \frac{1}{e}) \cdot n^{\Delta}(\mathbb{S})$, indicating that ESGR is a $\frac{1}{2} \cdot (1 - \frac{1}{e})$-approximation algorithm to SSMP. The theorem follows. □

Theorem 5. *The time complexity of ESGR is* $\min\{O(|V|^2), O(b \cdot |V|)\}$.

Proof. We prove this theorem in the Appendix B, available in the online supplemental material. □

To further boost the efficiency of ESGR, we propose two new pruning methods, named *Triangle Pruning (TP)* and *Subset Pruning (SP)*, which are able to remove redundant vertices from $\mathbb{V}$. Please refer to Appendix D, available in the online supplemental material, for the details.

6 Results of Evaluation Study

In the following, we first detail the setup of this evaluation study, then introduce *mixed effect model*, a statistical technique to quantify the difference between the control and experimental groups, to evaluate the long-term intervention effect quantitatively, and then presents the results of this evaluation study.

6.1 Setup

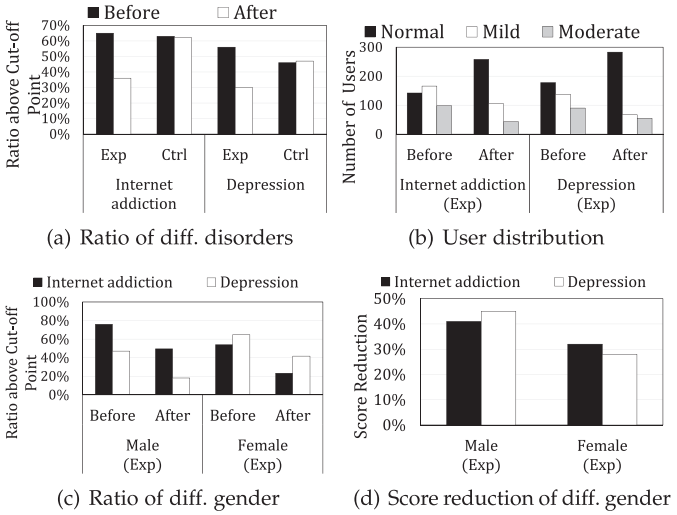
This evaluation study aims at illustrating the utility and feasibility of the proposed network intervention algorithm in real-world scenarios. We have recruited 1,020 volunteers, and after removing unqualified volunteers, there are 812 volunteers (referred to as *participants* hereafter) who completed this study. The study spanned 8 weeks, i.e., from June 2019 to August 2019. The recruited 812 volunteers include 408 males and 404 females, and their ages range from 20 to 35 years old. Most of the participants are university students and staffs in a national university in Taiwan, who form a social network with 812 vertices, 2,827 edges, and 1,348 4-SGs before the study begins. This 8-week study includes 8 weekly measurements of psychological outcomes among the 812 participants, and one additional pre-test outcome (measured before the study begins), resulting in 7,308 data points in this study. Two self-reported standard psychological questionnaires for *Internet Addiction* and *Depression* [35], [36] are adopted as the indicators of the health outcomes. The outcome of each questionnaire is an integer score to measure the severity of the disorder, and a higher score implies that the individual is suffering from more severe symptoms.

To evaluate the effectiveness of the network intervention recommended by the proposed ESGR algorithm, we randomly selected 406 participants who form 674 4-SGs as the experimental group (*Exp*), i.e., the network intervention is performed on this group, based on the recommendation of ESGR, to minimize the number of Δ-SGs. The other 406 participants, who form 670 of 4-SGs were considered the control group (*Ctrl*), i.e., no intervention is carried out on this group. It is worth noting that the suggestion for $\Delta = 4$ is merely for the current experimental group. Mental health professionals may suggest different Δ values for different scenarios and networks. In this evaluation study, we invited 9 mental health professionals to join the network intervention, who are from Taipei City Government Community Mental Health Center, National Taipei University of Nursing and Health Science, National Taiwan University, National Tsing Hua University, etc. In other words, $s = 9$ in this study. The mental health professionals suggested to set $b = 90$ by considering their time and workload. We employed our proposed ESGR algorithm to identify $b = 90$ users in the experimental group to add edges to the mental health professionals (i.e., adding 90 new edges) to reduce the number of 4-SGs in the intervened network.

Following the recommendations from the proposed ESGR algorithm with $s = 9$ and $b = 90$, the mental health professionals built 90 friendship links with the participants selected by ESGR (these 90 selected participants are referred to as *recipients*) and helped the recipients organize social activities for the rest of the participants in the experimental group. Those activities include online chatting and face-to-face hangouts. To validate the effectiveness of the network intervention, the mental health professionals only provided guidance and help to those recipients, but did not participate in the social activities directly. Also, the mental health professionals did not have any interaction with the participants other than the recipients.

6.2 Mixed Effect Model and Evaluation Criteria

American Psychological Association [37] provides general guidelines to systematically evaluate the efficacy of a



(a) Ratio of diff. disorders (b) User distribution

(c) Ratio of diff. gender (d) Score reduction of diff. gender

Fig. 6. Evaluation study results.

psychological intervention. The criteria for evaluating intervention suggest that attention should be paid to a number of issues, including the intervention goal, participant selection, and long-term consequences of the intervention.

Specifically, for *participant selection*, we randomly assigned our research subjects into the control group and the experimental group. Then, for *intervention goal*, we setup our goal to improve the mental well-being of the experimental group with network intervention. Therefore, during this experimental period of 8 weeks, each participant was asked to fill in two mental health questionnaires weekly, to let us understand the changes of each participant's mental status. Finally, for *long-term consequences of intervention*, we adopted the *mixed effect model* [38], [39], [40], a statistical technique to quantify the difference between the control and experimental groups, to evaluate the long-term intervention effect quantitatively.

The mixed effect model (also called *multilevel model*) allows researchers to systematically account for the dependency among repeated measurements in the same subjects and for the dependency among subjects nested in the same experimental groups [38], [39], [40]. The method is widely used in psychology and sociology, especially for psychological or network intervention for long-term behavioral change (e.g., [14], [41], [42]). Therefore, in our evaluation study, we adopt the mixed effect model, which incorporates the time variable and its interaction with intervention to allow us better describing the trajectory of changes in a subject's mental state over time and to delineate possible long-term effect of the intervention.

6.3 Results of Evaluation Study

In Fig. 6, the bars `before` and `after` refer to the average scores of the mental health outcomes before and after intervention, respectively.

In Fig. 6a, we first compare the ratios of the participants whose scores (scores of Internet addiction and depression from the standard psychological questionnaires) are above the cut-off points for each mental disorder, i.e., the ratios of participants who are considered to have at least mild or moderate symptoms of each type of mental disorder, before and after the network intervention organized by ESGR is performed. A larger ratio implies that more participants are suffering from mental disorders. The cut-off points are selected according to the suggestions of the psychological scales [35], [36] and the mental health professionals[4]. To better demonstrate the effectiveness of the network intervention organized by ESGR, we plot the results of both the experimental group (`Exp`) and control group (`Ctrl`). Please note that the control group was not involved in any kind of network intervention, i.e., there is no Δ-SG reduction for them.

The results in Fig. 6a manifest that the ratios drop significantly after the intervention for the experimental group, indicating that the mental well-being of a large number of participants indeed improves after intervention. In contrast, the ratios of the control group do not have any significant change. To further demonstrate the effectiveness of the proposed SSMP and ESGR, we also present the distributions of the participants with different severity levels before and after the network intervention organized by ESGR. In Fig. 6b, three types of bars, i.e., `Normal`, `Mild`, and `Moderate`, present the numbers of participants with different severity of symptoms for each mental disorder in the experimental group. Please note that there is no participant with severity level higher then *Moderate* in this study. Fig. 6b shows that the numbers of participants belonging to *Normal* increase and the numbers of participants belonging to *Moderate* decrease after intervention for both mental disorders. This indicates that many participants who originally had mild or moderate symptoms become better after the intervention organized with ESGR.

We then analyze the results based on the gender of the participants. Fig. 6c shows the ratios of the male and female participants in the experimental group whose scores are above the cut-off points for each mental disorder, and Fig. 6d presents the score reduction of the experimental group after intervention. The results indicate that after the intervention organized by ESGR, the ratios for both male and female participants, as well as the scores, drop significantly, indicating that the mental well-being of both female and male participants are significantly improved after intervention. Moreover, the reduction of the ratios and scores for males is more significant as compared to those for females. A possible reason behind this is that females are more likely to be depressed than males [43], [44]. Therefore, the reduction of the ratios for the female participants are not as much as that for males. However, the network intervention organized by ESGR still effectively help improve the female participants' well-being.

Furthermore, we also evaluate the effectiveness of the intervention recommended by ESGR with mixed effect modeling [38], [39], [40], a statistical technique to examine if the experimental and control groups are statistically different. Mixed effect modeling employs techniques similar to regressions. In the following, we first formulate the model to include important variables, and then fit the model with the data obtained in our evaluation study. The estimates (i.e., parameters) obtained after fitting the model can help

4. The cut-off points of Internet addiction and depression are 40 out of 100 points and 10 out of 45 points, respectively.

TABLE 3
Summary of Datasets

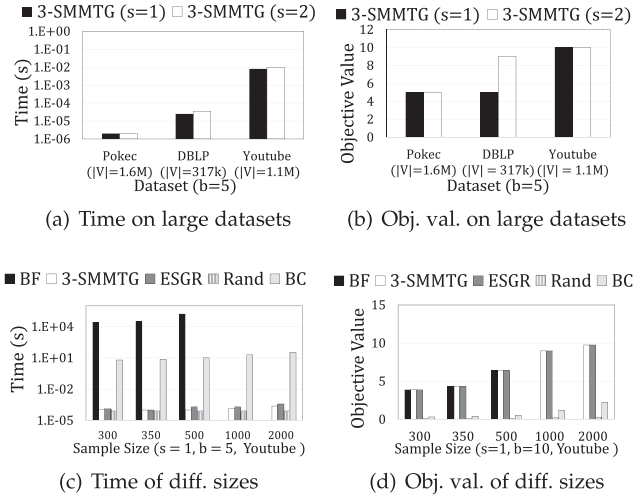
Dataset	$\|V\|$	$\|E\|$	CC	Diam
ego-FB	4K	44K	0.6	8
FB-GT	1.4K	14K	0.7	7
ego-Twitter	81K	1.7M	0.56	7
ego-Gplus	100K	13.6M	0.49	6
DBLP	317K	1M	0.63	21
Pokec	1.6M	30M	0.11	11
Youtube	1.1M	3M	2.7	24

us compare and understand the differences between the two groups. The detailed formulation of the mixed model and the results are presented in Appendix E, available in the online supplemental material. In summary, the model fitting results show that by employing ESGR for network intervention, the participants in the experimental group have significant improvements against Internet addiction and depression, as compared to the control group. This validates the problem formulation of SSMP and the algorithm design of ESGR proposed in this paper. Please refer to Appendix E, available in the online supplemental material, for the details.

7 Experimental Results

We conduct experiments on 7 real datasets to evaluate the proposed algorithms. First four datasets, *ego-FB* [45], *FB-GT* [24], *ego-Twitter* [45], and *ego-Gplus* [45], are social network datasets from Facebook, Twitter, and Google Plus. The ego-FB dataset contains 4K vertices and 44K edges, FB-GT contains 1.4K vertices and 14K edges, ego-Twitter contains 81K vertices and 1.7M edges, and ego-Gplus contains 100K vertices and 13.6M edges. Moreover, the fifth dataset, DBLP [46], is a co-author network with 317K vertices and 1M edges, and the sixth dataset, Pokec [47], is a social network with 1.6M vertices and 30M edges. Finally, the seventh dataset is the Youtube [46] social network with 1.1M vertices and 3M edges. The datasets are summarized in Table 3.

Since no algorithm has been proposed for SSMP, we compare our proposed 3-SMMTG and ESGR with three baseline algorithms: Brute-Force (BF), Random (RND), and BigClam (BC) [22]. BF finds the optimal solution of SSMP by enumerating all possible combinations of vertices satisfying the budget and seed constraints. RND randomly selects s vertices to form the set $\widehat{S}$. Then, RND treats $\widehat{S}$ as the b-Max Set and constructs the corresponding output graph with the steps similar to ESGR. That is, RND first constructs the set A with $\max_{u \in \widehat{S}} c_u$ vertices, and for each $v \in \widehat{S}$, it links v to each vertex in $\{a_1, \ldots, a_{c_v}\} \subseteq A$ and creates corresponding edges. BC is a large-scale community detection method. After finding all communities in the network with BC, we iteratively pick one vertex from each community randomly and add the selected vertex into the set $\widehat{S}$ until $|\widehat{S}| = s$. BC then treats $\widehat{S}$ as the b-Max Set and constructs the corresponding output graph with the steps similar to ESGR (as described above for RND). BC is implemented as a baseline because our purpose is to reduce the number of small dense subgroups in a social network. Since BC is able to find a set of dense communities, if we connect different members in different communities, we may have high chance to reduce the number of dense subgroups. The algorithms are implemented with C++ on an HP DL580 server with Quadcore Intel X5450 3.0 GHz CPUs and 1 TB RAM. The perprocessing step introduced earlier in this paper is performed offline.



(a) Time on large datasets (b) Obj. val. on large datasets

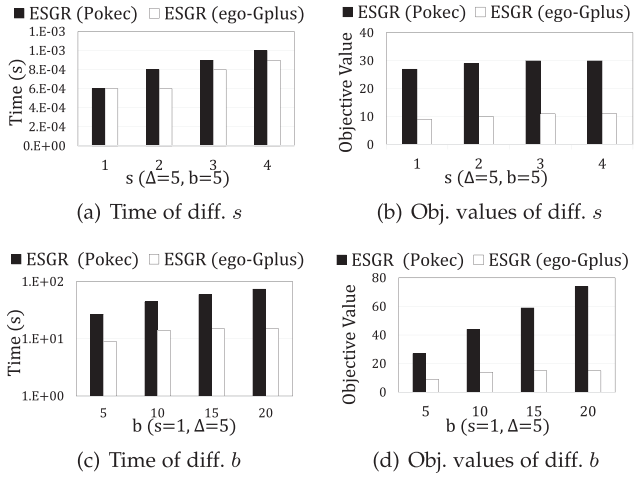
(c) Time of diff. sizes (d) Obj. val. of diff. sizes

Fig. 7. Evaluations of 3-SMMTG.

7.1 Evaluations of 3-SMMTG for SSMP with $\Delta = 3$

Sensitivity Tests on Large Graphs. Figs. 7a and 7b report the results of 3-SMMTG on Pokec ($|V| = 1.6M$), DBLP ($|V| = 317K$), and Youtube ($|V| = 1.1M$) to help understand the behavior of 3-SMMTG in different datasets. Figs. 7a and 7b compare the computation time and objective values in different datasets with different s, where s is the number of new vertices that can be added to the network. As s increases, the computation time and objective values increase because more vertices in V can be considered moving into $\mathbb{V}$. Youtube incurs the largest objective values as shown in Fig. 7b, because it has a smaller average degree than the other datasets, i.e., the average degree of Youtube is 5.4, whereas Pokec and DBLP have average degrees 37.5 and 6.3, respectively. As a consequence, more vertices remain unpruned in Youtube after the Cost Pruning, and more vertices imply a larger number of Δ-SGs in the graph. In this case, 3-SMMTG is able to achieve a larger objective value because it obtains the optimal solutions and thus is able to reduce the maximum number of Δ-SGs.

Comparisons with Other Baselines. We also compare the proposed 3-SMMTG with other baseline approaches. Since the brute-force approach, BF, is unable to return a solution in 24 hours when the graph contains more than 500 vertices, we randomly sample the Youtube dataset to generate small networks with different sizes. Figs. 7c and 7d compare 3-SMMTG with BF, RND, and ESGR, i.e., the approximation algorithm for the general SSMP, with different sample sizes. Fig. 7c shows that even on very small graphs, BF still incurs unacceptable computation time to find the optimal solution. In contrast, 3-SMMTG and ESGR are very efficient because they are equipped with effective pruning strategies to avoid examining redundant vertices. In particular, 3-SMMTG is able to obtain the optimal solution very efficiently, and thus it has a smaller computation time as compared to ESGR.



(a) Time of diff. s (b) Obj. values of diff. s

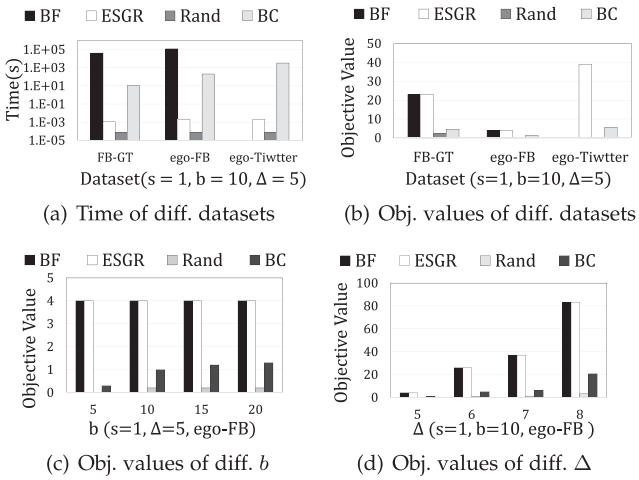
(c) Time of diff. b (d) Obj. values of diff. b

Fig. 8. Sensitivity tests of ESGR on large datasets.

Fig. 7d demonstrates that 3-SMMTG and ESGR are able to obtain high-quality solutions. As proved in Section 4, the proposed 3-SMMTG can find the optimal solution and thus 3-SMMTG has the objective values exactly the same as those of BF when the sample size does not exceed 500 (BF cannot return a solution within 24 hours when the sample size is larger than 500). Moreover, although ESGR is an approximation algorithm, it is able to find the optimal solutions in all these test cases because ESGR extracts the set S exactly the same as that of 3-SMMTG. In contrast, RND and BC perform poorly because they cannot effectively reduce the number of Δ-SGs.

7.2 Evaluations of ESGR for General SSMP

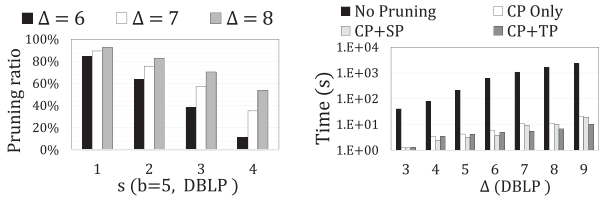
Sensitivity Tests on Large Graphs. Fig. 8 shows the results of the proposed ESGR on different datasets, e.g., Pokec ($|V| = 1.6M$) and ego-Gplus ($|V| = 100K$). Fig. 8a shows that as s increases, the computation time increases as well because fewer vertices are pruned by the Cost Pruning. As shown in Fig. 8b, when s becomes larger, the objective values also increase. This is because for a larger s, there are more different choices of vertices for maximizing the objective function. We also evaluate ESGR with different b in Figs. 8c and 8d. When b increases, there are more edges that can be connected from the newly added vertices and the original vertices. ESGR in this case can eliminate more Δ-SGs and increase the objective value. Since the average degree of Pokec (37.5) is much smaller than that of ego-Gplus (272), ESGR obtains larger objective values in Pokec. The reason is similar to that for Fig. 7b in Section 7.1.



(a) Time of diff. datasets (b) Obj. values of diff. datasets

(c) Obj. values of diff. b (d) Obj. values of diff. Δ

Fig. 9. Comparisons with baselines.



(a) Performance of Cost Pruning (b) Pruning strategies with different Δ

Fig. 10. Comparisons of pruning strategies.

Comparisons with Other Baselines. We compare the proposed ESGR with other baseline approaches (BF, BC, RND) in Fig. 9. Since BF does not scale up to large social networks, we compare ESGR with these baselines on small real datasets. As shown in Fig. 9a, even for the small networks, BF still incurs very large computation time to find the optimal solution. In contrast, the Cost Pruning equipped by ESGR effectively avoids the examinations of redundant vertices. Since BF cannot obtain a solution in 24 hours in ego-Twitter, the results of BF in ego-Twitter in Figs. 9a and 9b are not plotted.

Fig. 9b compares the solution quality on different datasets. ESGR obtains high-quality solutions. Even if ESGR is a $\frac{1}{2}(1-\frac{1}{e})$-approximation algorithm, ESGR is still able to obtain the optimal solutions in FB-GT and ego-FB (BF does not return a solution for ego-Tiwtter, and whether ESGR obtains the optimal solution in ego-Twitter is unknown). In contrast, RND and BC both perform poorly. Figs. 9c and 9d report the results of ESGR and other baselines with different parameters b and Δ on ego-FB. Fig. 9c indicates that ESGR outperforms Rand and BC for different b. In Fig. 9d, the objective values of ESGR increase as Δ increases because more Δ-SGs appears with a larger Δ.

Effectiveness of Pruning Strategies. Fig. 10a reports the performance of Cost Pruning (CP) in DBLP. The *pruning ratio* is the ratio calculated as the number of pruned vertices divided by the input graph size. A larger pruning ratio indicates the pruning strategy is more effective. We observe that when s is smaller, CP is able to prune more redundant vertices. The reason is that for a vertex v, its cost $c_v = \Delta - deg_G(v) + 1$. The larger the Δ is, the larger the c_v will be. Since the Cost Pruning prunes the vertices with costs greater than s, a larger Δ implies that more vertices can be pruned.

Fig. 10b demonstrates the effectiveness of different pruning strategies. ESGR takes much more time when no pruning is employed, indicating that the proposed pruning strategies are very effective. As SP is more powerful than TP, it is able to prune more vertices than TP does. However, the running time of SP increases significantly as Δ increases. Fig. 10b shows that when $\Delta \leq 6$, CP+SP outperforms CP +TP. However, when $\Delta > 6$, the extra computation time of SP is more significant and thus CP+TP incurs a smaller computation time.

8 Conclusion

In this paper, we first propose the new notion of Δ-SGs to quantify the small dense subgroups in social networks. Then, we formulate a new research problem, SSMP, to reduce the number of small dense subgroups in the network. For the special case of SSMP with $\Delta = 3$, we propose a linear-time algorithm 3-SMMTG to obtain the optimal solution. For the general SSMP, we propose algorithm ESGR, which is a $\frac{1}{2}(1 - \frac{1}{e})$-approximation algorithm. Our 8-week evaluation study with 812 participants validates the proposed SSMP and ESGR, i.e., showing that the participants with the network intervention recommended by ESGR have significant improvements on Internet addiction and depression, as compared to those individuals without any intervention. Moreover, experimental results on real datasets also show that the proposed algorithms outperform the baselines in both efficiency and solution quality. In our future work, we will explore potential approaches to improve the performance of ESGR and extend SSMP to consider more factors for network intervention.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan, under MOST 108-2636-E-007-009- (MOST Young Scholar Fellowship Columbus Program), 108-2218-E-468-002-, and 107-2218-E-002-010-.

References

[1] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 43:1–43:35, 2013.
[2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge, U.K., Cambridge Univ. Press, 1994.
[3] U. Feige, D. Peleg, and G. Kortsarz, "The Dense k-Subgraph Problem," *Algorithmica*, vol. 29, pp. 410–421, 2001.
[4] R. J. Mokken, "Cliques, Clubs and Clans," *Quality Quantity: Int. J. Methodology*, vol. 13, pp. 161–173, 1979.
[5] F. Hao, D.-S. Park, Z. Pei, H. Lee, and Y.-S. Jeong, "Identifying the social-balanced densest subgraph from signed social networks," *J. Supercomputing*, vol. 72, no. 7, pp. 2782–2795, 2016.
[6] S. Meeks and S. A. Murrell, "Service providers in the social networks of clients with severe mental illness," *Schizophrenia Bulletin*, vol. 20, no. 2, pp. 399–406, 1994.
[7] D. Wasylenki, S. James, C. Clark, J. Lewis, P. Goering, and L. Gillies, "Clinical issues in social network therapy for clients with schizophrenia," *Community Mental Health J.*, vol. 28, no. 5, pp. 427–440, 1992.
[8] D. L. Cutler, E. Tatum, and J. H. Shore, "A comparison of schizophrenic patients in different community support treatment approaches," *Community Mental Health J.*, vol. 23, no. 2, pp. 103–113, 1987.
[9] L. Maguire, *Understanding Social Networks*. Sage Publications, Inc, vol. 32, 1983.
[10] I. Kawachi and L. F. Berkman, "Social ties and mental health," *J. Urban Health*, vol. 78, no. 3, pp. 458–467, 2001.
[11] E. Pattison, D. Defrancisco, P. Wood, H. Frazier, and J. Crowder, "A psychosocial kinship model for family therapy," *American J. Psychiatry*, vol. 132, no. 12, pp. 1246–1251, 1975.
[12] K. L. Fiori and J. Jager, "The impact of social support networks on mental and physical health in the transition to older adulthood: A longitudinal, pattern-centered approach," *Int. J. Behavioral Development*, vol. 36, no. 2, pp. 117–129, 2012.
[13] R. M. Pinto, "Using social network interventions to improve mentally ill clients' well-being," *Clinical Social Work J.*, vol. 34, no. 1, pp. 83–100, 2006.
[14] T. A. Pickering et al., "Diffusion of a peer-led suicide preventive intervention through school-based student peer and adult networks," *Frontiers Psychiatry*, vol. 9, 2018, Art. no. 598.
[15] E. J. Van Den Oord and R. Van Rossem, "Differences in first graders' school adjustment: The role of classroom characteristics and social structure of the group," *J. School Psychology*, vol. 40, no. 5, pp. 371–394, 2002.
[16] D. Kamps et al., "A comprehensive peer network intervention to improve social communication of children with autism spectrum disorders: A randomized trial in kindergarten and first grade," *J. Autism Developmental Disorders*, vol. 45, no. 6, pp. 1809–1824, 2015.
[17] L. R. Derogatis, *SCL-90-R : Administration, Scoring of Procedures Manual-II for the R(evised) Version and other Instruments of the Psychopathology Rating Scale Series*. Towson, Md.: Clinical Psychometric Research Inc., 1992.
[18] K. L. Fiori, T. C. Antonucci, and K. S. Cortina, "Social network typologies and mental health among older adults," *J. Gerontology: Series B*, vol. 61, no. 1, pp. P25–P32, 2006.
[19] P. J. Flores and J. Georgi, *Substance Abuse Treatment: Group Therapy*. Washington, DC, USA: U.S. Dept. Human Health Services, 2005.
[20] C. Salmivalli, A. Huttunen, and K. M. J. Lagerspetz, "Peer networks and bullying in schools," *Scandinavian J. Psychology*, vol. 38, no. 4, pp. 305–312, 1997.
[21] J. Hastad, "Clique is Hard to Approximate within $n^{1-\epsilon}$," *Acta Mathematica*, vol. 182, pp. 105–142, 1999.
[22] J. Yang and J. Leskovec, "Overlapping community detection at scale: A nonnegative matrix factorization approach," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 587–596.
[23] Q. Zhu, H. Hu, C. Xu, J. Xu, and W.-C. Lee, "Geo-social group queries with minimum acquaintance constraints," *VLDB J.*, vol. 26, no. 5, pp. 709–727, 2017.
[24] H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, "Willingness optimization for social group activity," *Proc. VLDB Endowment*, vol. 7, no. 4, pp. 253–264, 2013.
[25] S. Lin, L. Faust, P. Robles-Granda, T. Kajdanowicz, and N. V. Chawla, "Social network structure is predictive of health and wellness," *PloS One*, vol. 14, no. 6, 2019, Art. no. e0217264.
[26] A. Dhand, D. A. Luke, C. E. Lang, and J.-M. Lee, "Social networks and neurological illness," *Nat. Rev. Neurol*, vol. 12, no. 10, 2016, Art. no. 605.
[27] H. Lin et al., "Detecting stress based on social interactions in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1820–1833, Sep. 2017.
[28] B. S. Fraga, A. P. C. da Silva, and F. Murai, "Online social networks in health care: A study of mental disorders on reddit," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2018, pp. 568–573.
[29] T. Wang, M. Brede, A. Ianni, and E. Mentzakis, "Social interactions in online eating disorder communities: A network perspective," *PloS One*, vol. 13, no. 7, 2018, Art. no. e0200800.
[30] H.-H. Shuai et al., "A comprehensive study on social network mental disorders detection via online social media mining," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1212–1225, Jul. 2018.
[31] K. O'Leary, S. M. Schueller, J. O. Wobbrock, and W. Pratt, "Suddenly, we got to become therapists for each other: Designing peer support chats for mental health," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, Art. no. 331.
[32] E. L. Murnane, T. G. Walker, B. Tench, S. Voida, and J. Snyder, "Personal informatics in interpersonal contexts: towards the design of technology that supports the social ecologies of long-term mental health management," *Proc. ACM Human-Comput. Interaction*, vol. 2, 2018, Art. no. 127.
[33] H.-H. Shuai, Y.-C. Lien, D.-N. Yang, Y.-F. Lan, W.-C. Lee, and P. S. Yu, "Newsfeed filtering and dissemination for behavioral therapy on social network addictions," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Management*, 2018, pp. 597–606.
[34] B. Wilder, H. C. Ou, K. de la Haye, and M. Tambe, "Optimizing network structure for preventative health," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst.*, 2018, pp. 841–849.
[35] K. S. Young, *Caught in the Net: How to Recognize the Signs of Internet Addiction–and a Winning Strategy for Recovery*. Hoboken, NJ, USA: Wiley, 1998.
[36] D. D. Burns, *Feeling Good: The New Mood Therapy*. New York, NY, USA: AvonBooks, 1999.
[37] D. Hollon et al., "Criteria for evaluating treatment guidelines," *American Psychologist*, vol. 57, no. 12, pp. 1052–1059, 2002.
[38] T. A. Snijders, *Multilevel Analysis*. Berlin, Germany: Springer, 2011.
[39] J. J. Hox, M. Moerbeek, and R. Van de Schoot, *Multilevel Analysis: Techniques and Applications*. Abingdon, U.K.: Routledge, 2017.
[40] A. Gałecki and T. Burzykowski, *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Berlin, Germany: Springer, 2013.

[41] D. Freeman et al., "The effects of improving sleep on mental health (oasis): A randomised controlled trial with mediation analysis," *Lancet Psychiatry*, vol. 4, no. 10, pp. 749–758, 2017.
[42] K. Weinger, E. A. Beverly, Y. Lee, L. Sitnokov, O. P. Ganda, and A. E. Caballero, "The effect of a structured behavioral intervention on poorly controlled diabetes: a randomized controlled trial," *Archives Internal Medicine*, vol. 171, no. 22, pp. 1990–1999, 2011.
[43] P. Albert, "Why is depression more prevalent in women?" *JPsychiatry Neuroscience*, vol. 40, pp. 219–221, 2015.
[44] F. Bahrami and N. Yousefi, "Females are more anxious than males: A metacognitive perspective," *Iranian J. Psychiatry Behavioral Sci.*, vol. 5, pp. 83–90, 2011.
[45] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 539–547.
[46] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 745–754.
[47] L. Takac and M. Zabovsky, "Data analysis in public social networks," in *Proc. Int. Sci. Conf. Int. Workshop Present Day Trends Innovations*, 2012, pp. 1–6.

Bay-Yuan Hsu received the MS degree from the Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan, in 2012, and the PhD degree from the Department of Computer Science, University of California at Santa Barbara in 2019. His research interests include bioinformatics, big data, and social network analytics.

Chih-Ya Shen received the PhD degree from the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, in 2013. He is an assistant professor with the Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan. His research interests include big data and social network analytics, query processing, and mobile computing. He received the MOST Young Scholar Fellowship (MOST Columbus Program) from the Ministry of Science and Technology, Taiwan, the NTHU New Faculty Research Award from National Tsing Hua University, and the K. T. Li Distinguished Young Scholar Award from ACM Taipei/Taiwan Chapter.

Xifeng Yan received the PhD degree in computer science from the University of Illinois at Urbana-Champaign, United States in 2006 and was a research staff member at the IBM T. J. Watson Research Center between 2006 and 2008. He is a professor with the University of California at Santa Barbara, holding the Venkatesh Narayanamurti Chair of Computer Science. His research interests include modeling, managing, and mining data, especially graph and text data. He received NSF CAREER Award, IBM Invention Achievement Award, ACM-SIGMOD Dissertation Runner-Up Award, and IEEE ICDM 10-year Highest Impact Paper Award.