This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Multiplex Influence Maximization in Online Social Networks With Heterogeneous Diffusion Models

Alan Kuhnle, *Student Member, IEEE*, Md Abdul Alim, Xiang Li,
Huiling Zhang, and My T. Thai, *Member, IEEE*

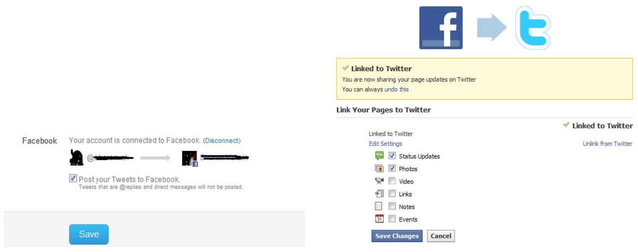
***Abstract*—Motivated by online social networks that are linked together through overlapping users, we study the influence maximization problem on a multiplex, with each layer endowed with its own model of influence diffusion. This problem is a novel version of the influence maximization problem that necessitates new analysis incorporating the type of propagation on each layer of the multiplex. We identify a new property, generalized deterministic submodular, which when satisfied by the propagation in each layer, ensures that the propagation on the multiplex overall is submodular–for this case, we formulate influential seed finder (ISF), the greedy algorithm with approximation ratio $(1-1/e)$. Since the size of a multiplex comprising multiple OSNs may encompass billions of users, we formulate an algorithm knapsack seeding of network (KSN) that runs on each layer of the multiplex in parallel. KSN takes an α-approximation algorithm A for the influence maximization problem on a single network as input, and has approximation ratio $((1-\epsilon)\alpha)/((o+1)k)$ for arbitrary $\epsilon > 0$, o is the number of overlapping users, and k is the number of layers in the multiplex. Experiments on real and synthesized multiplexes validate the efficacy of the proposed algorithms for the problem of influence maximization in the heterogeneous multiplex. Implementations of ISF and KSN are available at http://www.alankuhnle.com/papers/mim/mim.html.**

***Index Terms*—Approximation algorithms, heterogeneous networks.**

I. Introduction

THE rapid growth of large online social networks (OSNs) such as Facebook, Google+, and Twitter has enabled them to become thriving places for viral marketing in recent years. People are increasingly engaged in OSNs: 62% of adults worldwide use social media and spend 22% of online time on social networks on average [1]. Much like real-world social networks, information spreading in OSNs has viral properties, creating an excellent medium for marketing. Due to the impact of this effect on the popularity of new products, OSNs have rapidly become an attractive channel for raising awareness of new products or brands. In this context, an important problem is how to find the best set of seed users who can influence the most other users.

Increasingly, users engage in more than one OSN; they connect their accounts across multiple networks, such that posts in one network are simultaneously posted in other networks. In Fig. 1, we show the process of connecting a Facebook and Twitter account, which allows automatically posting on Facebook when a new tweet is sent, and vice versa. As a consequence, the propagation of information can cross from one OSN to another through these overlapping users.



Facebook Your account is connected to Facebook. (Disconnect)

@[redacted] [redacted]

☑ Post your Tweets to Facebook.
Tweets that are @replies and direct messages will not be posted.

(a)

✓ Linked to Twitter
You are now sharing your page updates on Twitter
You can always undo this

Link Your Pages to Twitter

✓ Linked to Twitter

Linked to Twitter
Edit Settings

Unlink from Twitter

☑ Status Updates
☑ Photos
□ Video
□ Links
□ Notes
□ Events

(b)

Fig. 1. Process of becoming an overlapping user of Twitter and Facebook. (a) Autopost from Twitter to Facebook. (b) Autopost from Facebook to Twitter.

The influence propagation in each OSN will be particular to that network; for example, usage patterns for Twitter and Facebook are quite different. Moreover, even different cascades in the same social network may be better explained by different models of influence propagation [2]. Thus, overlapping users connect OSNs together into a multiplex structure of OSNs, comprising multiple OSNs linked together through overlapping users, where each OSN has different local propagation. In this paper, we study the multiplex influence maximization (MIM) problem, to pick the most influential seed nodes, on a multiplex of OSNs as described above. Several natural questions can arise.

1) What conditions on the propagation in each layer OSN are sufficient for the overall multiplex propagation to have the submodularity property, which is important for approximation algorithms?
2) Can existing methods for single OSNs be utilized within a solution to the MIM problem?
3) What role do overlapping users play in the influence propagation on a multiplex of OSNs?

From a computational perspective, a multiplex consisting of multiple OSN layers may be very large, comprising billions of users in each layer.

To demonstrate how propagation in a multiplex differs from propagation in a single layer, consider the toy multiplex shown

Manuscript received September 18, 2017; revised January 19, 2018; accepted February 27, 2018. This work was supported in part by the NSF Grant CNS-1443905 and in part by DTRA grant HDTRA1-14-1-0055. *(Corresponding author: Alan Kuhnle.)*

The authors are with the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32603 USA (e-mail: kuhnle@ufl.edu; mythai@ufl.edu).

Digital Object Identifier 10.1109/TCSS.2018.2813262

2329-924X © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

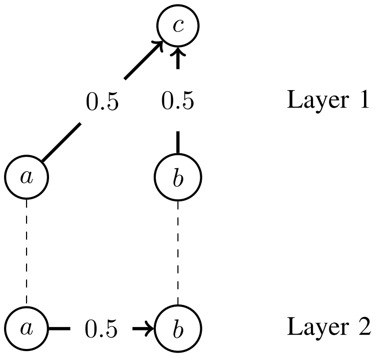


Fig. 2. Two-layer multiplex exemplifying how propagation in a multiplex will deviate from the propagation in a single layer.

-	LinkedIn	Facebook	Twitter	MySpace
LinkedIn	-	12%	21%	6%
Facebook	82%	-	91%	57%
Twitter	31%	20%	-	17%
MySpace	36%	49%	70%	-

Fig. 3. Percentage of overlapping users between major OSNs in 2009 [8]. The table is read as follows: $(x, y)\%$ of users in OSN y also use OSN x, where x is the row, y is the column.

in Fig. 2, where a, b are overlapping users in both layers, and c is only present in Layer 1. Let Layer 1 has a fixed threshold model, with the threshold θ_c of c equal to 1. Thus, c becomes activated if a, b are both activated. Let Layer 2 has the independent cascade (IC) model. Then, seeding vertex a will result in both b, c having a chance of becoming activated, as a may activate b according to the IC model in Layer 2. If b becomes activated, then a, b will together activate c in Layer 1. Finally, observe that the activation of the two layers cannot be incorporated into a single layer network following either the IC model or the fixed threshold model.

For influence maximization problem on a single layer network with propagation according to a single model, such as IC, approximation algorithms have been developed and optimized [3]–[7]. However, since these algorithms only consider a single model of influence propogation, they are not directly suitable for MIM, where each OSN has a different model of propagation. As shown in Fig. 3, the fraction of overlapping users is considerable.

Our main contributions are summarized as follows.

1) We define the generalized deterministic submodular (GDS) property, which if satisfied by each layer implies overall propagation on the multiplex is submodular. Submodularity allows the formulation of a greedy $(1 - 1/e)$-approximation algorithm–influential seed finder (ISF)–for MIM.
2) We provide an approximation algorithm KSNs, which is parallelizable by layer of the multiplex and utilizes an algorithm $\mathcal{A}$ for the single-layer influence maximization problem together with a knapsack-based approach to create a solution for MIM, thereby taking advantage of previous optimizations [4]–[7] for the homogeneous, single layer case. If the utilized algorithm $\mathcal{A}$ has approximation factor α, then KSN has approximation ratio
$$\frac{(1-\epsilon)\alpha}{(o+1)k}$$
where o is the number of overlapping users in the multiplex, k is the number of layers in the multiplex, and $\epsilon > 0$ is arbitrary.
3) Experimental evaluation of all algorithms on a variety of multiplexes, both synthesized and from traces of real multiplexes of OSNs, validates the effectiveness of the two approximation algorithms.

Overlapping users can be identified in real networks using, for example, methods in [9] and [10]; methods of identification is not a focus of this paper.

The rest of the paper is organized as follows. In Section II, we provide technical definition for model of influence propagation. Also, we present the influence propagation model in multiplex networks in terms of its component layers and define the problem. We then outline our proposed algorithms for solving the MIM problem along with the inapproximability proof for a class of models in Section III. Section IV shows the experimental results on the performance of different algorithms. In Section V, we discuss related work on influence propagation and finally, Section VI concludes this paper.

II. Model Representation and Problem Formulation

A. Influence Propagation Models

Intuitively, the idea of a model of influence propagation in a network is clear; it is a way by which nodes can be activated or influenced given a set of seed nodes. Kempe *et al.* [11] studied a variety of models in their seminal work on influence progation on a graph, including the IC and linear threshold (LT) models. For completeness, we briefly describe these two models. An instance of influence propagation on a graph G follows the IC model if a weight can be assigned to each edge such that the propagation probabilities can be computed as follows: once a node u first becomes active, it is given a single chance to activate each currently inactive neighbor v with probability proportional to the weight of the edge (u, v). In the LT model, each network user u has an associated threshold $\theta(u)$ chosen uniformly from $[0, 1]$, which determines how much influence (the sum of the weights of incoming edges) is required to activate u. u becomes active if the total influence from its direct neighbors exceeds the threshold $\theta(u)$.

In this paper, since we allow each layer of a multiplex to have a different model of influence propagation, we need a technical definition for this concept.

Definition 1 (Model of Influence Propagation): A model of influence propagation σ on a graph $G = (V, E)$ is a function P that assigns, for each $A \subset V$, and for each $S \subset V$, a probability
$$P(S|A) = P(S \text{ is final activated set } |A \text{ is seed set }) \in [0, 1],$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

satisfying

$$\begin{aligned}&(1)\ \text{if } B \cap A \subsetneq A P(B|A) = 0,\\ &(2) \sum_{S:S\subset V} P(S|A) = 1.\end{aligned}$$

(1) simply states that seed nodes may not become unactivated, and (2) ensures that we have a probability distribution.

The expected number of activated nodes given a seed set A is denoted $\sigma(A)$, and

$$\sigma(A) = \sum_{S:S\subset V} P(S|A) \cdot |S|.$$

A model σ is called *deterministic* if for each $A \subset V$, there exists F_A such that $P(F_A|A) = 1$; intuitively, deterministic means that there is no probability in the model of diffusion since the final set activated is uniquely determined by the seed set. If σ is deterministic, $\sigma(A) = |F_A|$; we abuse notation and also use $\sigma(A) = F_A$, the final set itself. This allows convenient specification of the set $T = \sigma(\tau(A))$, for example, where both σ, τ are deterministic models, and T is the final activated set generated by using the final set of A under τ as the seed set for σ.

Many models of information propagation discussed in the literature satisfy the submodularity property, that σ satisfies

$$\sigma(A) + \sigma(B) \geq \sigma(A \cup B) + \sigma(A \cap B)$$

for all $A, B \subset V$. Submodularity is important since it guarantees that a greedy approach to the influence maximization prolem will have an approximation ratio [12]. We now define a property that is stronger than submodularity.

Definition 2 (Generalized Deterministic Submodular): Let σ be a model of influence propagation. σ satisfies the GDS property if the expected number of activations, given seed set A, can be written

$$\sigma(A) = \sum_{j=1}^{s} p_j \sigma_j(A)$$

where each σ_j, $j \in \{1, \ldots, s\}$ is a deterministic, submodular model of influence propagation, and $p_j \in [0, 1]$, $\sum_{j=1}^{s} p_j = 1$.

Lemma 1: Let σ be a model of influence propagation. If σ satisfies GDS, then σ is submodular.

Proof: Let A be an arbitrary seed set. Since σ satisfies GDS, $\sigma(A) = \sum_{i=1}^{s} p_j \sigma_j(A)$, where $\sigma_j(A)$ is expected activation of deterministic and submodular model σ_j. Hence, the expected activation function σ is a nonnegative linear combination of submodular functions, thus σ is submodular. □

Examples of models in the literature that satisfy submodularity include IC, LT, Asynchronous Independent Cascade, asynchronous linear threshold [2], independent cascade model for endogenous competition, homogeneous competitive independent cascade model, and K-LT competitive diffusion model, as well as others [11], [14]. In all cases, the submodularity of these models has been shown by considering instances where some edges are live and some are blocked–each such instance corresponds to a deterministic submodular model. Thus, all of these proofs show submodularity by showing the stronger property GDS and using Lemma 1. Another example of a model that satisfies GDS is the conformity and context-aware cascade model [14].

In Section III, we show (Theorem 1) that influence propagation on a multiplex satisfies GDS if the propagation on each layer network satisfies GDS; hence, if the propagation on each layer satisfies GDS, the propagation in the multiplex is submodular.

B. Notations and Multiplex Model

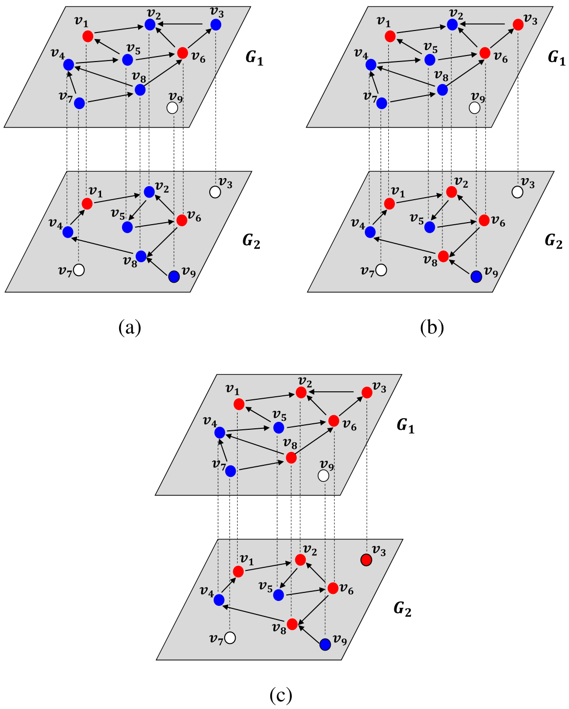
A social network can be modeled as a directed graph $G = (V, E)$. The vertex set V represents the participation of users in the social network, and the edge set represents the connections among network users. These connections model friendships or relationships.

Definition 3 (Heterogeneous Multiplex): A multiplex of OSNs is a list $\mathscr{G} = \{(G_i, \sigma_i) : i \in \{1, \ldots, k\}\}$ with $G_i = (V_i, E_i)$ a directed graph representing an OSN and influence model σ_i, representing the model of influence propagation in G_i. If a user belongs to more than one OSN, an interlayer edge is added between the pair of nodes, one in each OSN, representing this user. Such a user is termed *overlapping user*; we will denote the set of overlapping users by O. We will denote the set of all users in the multiplex by $V = \bigcup_{i=1}^{k} V_i$.

The influence propagation model σ on the multiplex is defined in the following way. If an overlapping vertex v is activated in one graph G_i, then, deterministically its adjacent interlayer copies become activated in all OSNs; propagation occurs in each graph G_i according to its propagation model σ_i. Fig. 4 shows an example of the definition of σ, in a multiplex $\mathscr{G} = \{(G_1, \sigma_1), (G_2, \sigma_2)\}$ with two layers. Here σ_1 and σ_2 are simply deterministic models of activation following the directed edges in the layers. Initially, the activated set is $\{v_1, v_6\}$ in Fig. 4(a), shown by red nodes. In Fig. 4(b), the activation has propagated according to σ_1, σ_2 in G_1, G_2, respectively, activating in addition set $\{v_2, v_3, v_8\}$. Next, in Fig. 4(c), the propagation proceeds between the two layers via overlapping nodes, whereupon it may continue in each layer G_i according to σ_i. Propagation ceases when no new nodes are activated in any layer. In addition, Fig. 4 demonstrates how without loss of generality, we may consider all nodes to be overlapping by adding absent nodes as isolated nodes (the white nodes). In Section III-A1, we make use of this fact by considering all layers to have the same nodes–in the rest of the paper, we consider overlapping users to be nontrivial; i.e., there are no isolated nodes in any layer. We refer to the users that actively participate in multiple networks (i.e., are nonisolated) as overlapping users, which may be identified in real networks using, for example, methods in [9] and [10].

The expected number of activations in the multiplex given seed set $A \subset V$ is denoted $\sigma(A)$, in addition to denoting the model defined above as σ; we do not count more than one copy of overlapping nodes toward $\sigma(A)$. Each graph G_i is referred to as a *layer network* of the multiplex G. We refer to a graph $G = (V, E)$ that is not part of a multiplex as a single network,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a)

(b)

(c)

Fig. 4. (a)–(c) Example of influence propagation σ in the multiplex structure. Red nodes are activated, blue and white are unactivated nodes. The white color indicates a node added to the layer in order that each network has the same vertex set V. In (a), we have seed set conserting of v_1 and v_6. In (b), propagation occurs in each layer according to that layers model, and in (c), nodes activated in one layer are activated in all, whence the propagation may continue.

to contrast with multiplex network, and we refer to propagation occuring in a single network according to a single influence propagation model as homogeneous propagation, to contrast with the heterogeneous propagation in a multiplex with more than one layer and propagation model in each layer.

C. Problem Definition

We now consider the problem of maximizing the influence of a seed set of given size in a multiplex network. Formally,

Definition 4 (Heterogeneous Multiplex Influence Maximization): Given a multiplex network $\mathcal{G} = (G_1, \sigma_1), \ldots, (G_k, \sigma_k)$ with k layer networks, influence propagation model σ on $\mathcal{G}$, as defined above in terms of σ_i, and positive integer l, find a set $S \subset V$ of size at most l so as to maximize the expected number of active users $\sigma(S)$. An instance of this problem will be denoted $(\mathcal{G}, k, l, \sigma)$.

III. Approximations of MIM

Since influence maximization on a single network is a special case of influence maximization on a multiplex, MIM is NP-complete. In this section, we first prove that the propagation σ on a multiplex is submodular if the propagation on each layer satisfies GDS and formulate a greedy algorithm to maximize expected influence. If each layer satisfies GDS, σ is submodular, and thus our greedy algorithm has approximation ratio $1 - 1/e$. Finally, we consider approaches to approximate MIM by approximating influence maximization on each layer separately and effectively combining the result into a feasible solution for MIM, which leads to a scalable approximation algorithm KSN with ratio depending on number of overlapping users o, number of layers k in the multiplex, the ratio α for the approximation used on the homogeneous layers, and an arbitrary $\epsilon > 0$; the ratio of KSN is $((1-\epsilon)\alpha)/((o+1)k)$.

A. Greedy Approach

Let $\mathcal{G} = (G_i, \sigma_i)_{i=1}^k$ be a multiplex with propagation model σ. We prove that σ is submodular for the case that each σ_i satisfies generalized deterministic submodularity (GDS). Thus, the greedy algorithm, which we detail in this section, achieves a $(1 - 1/e)$ ratio when propagation in each layer σ_i satisfies GDS.

1) Submodularity: Without loss of generality, we may consider that the sets V_i are the same, that is, $V_i = V$ for all i and some set V: if a vertex $v \in G_i$ does not exist in some G_j, simply add it to G_j as an isolated vertex. Thus, in this section only, we consider $V_i = V$ for all i. Recall that instead of counting the activation of all k copies of node $u \in V$, we count only a single copy as activated. The expected number of activations in the multiplex given seed set $A \subset V$ is denoted $\sigma(A)$; again, we do not count more than one copy of $u \in V$ toward $\sigma(A)$.

We will first consider a simpler case: when the propagation of each G_i is deterministic and submodular.

a) Deterministic case: In this section, let the propagation σ_i of each G_i be deterministic and submodular. Recall the definition of the multiplex influence propagation σ. Given a seed set $S \subset V$, the set of nodes in the multiplex that are activated after the propagation finishes will be denoted by $\tau(S)$; the nodes activated if propagation is restricted to only G_i will be denoted $\tau_i(S)$. Notice that $|\tau(S)| = \sigma(S)$ and $|\tau_i(S)| = \sigma_i(S)$.

Lemma 2: Let $S \subset V$. Then $\tau_i(\tau(S)) = \tau(S)$ for all i.

Proof: This follows from definition of propagation in multiplex. If propagation has resulted in set $\tau(S)$, then propagation cannot proceed further in any layer network, for otherwise propagation in $\mathcal{G}$ would not have terminated with $\tau(S)$. A visualization of an example of multiplex propagation is shown in Fig. 4 (a)–(c). □

Lemma 3: Let $S, T \subset V$.

$$\tau(S) \cup \tau(T) = \tau_i(\tau(S) \cup \tau(T))$$

for all i.

Proof: We have

$$\sigma_i(\tau(S) \cup \tau(T)) + \sigma_i(\tau(S) \cap \tau(T)) \leq \sigma_i(\tau(S)) + \sigma_i(\tau(T))$$

by submodularity of σ_i, so

$$\begin{aligned}\sigma_i(\tau(S) \cup \tau(T)) &\leq |\tau(S)| + |\tau(T)| - |\tau(S) \cap \tau(T)| \\ &= |\tau(S) \cup \tau(T)|\end{aligned}$$

by Lemma 2 and since $\sigma_i(\tau(S) \cap \tau(T)) \geq |\tau(S) \cap \tau(T)|$. Hence,

$$\tau_i(\tau(S) \cup \tau(T)) = \tau(S) \cup \tau(T).$$

□

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Lemma 4: Let $S, T \subset V$.

1) $\tau(S \cup T) = \tau(S) \cup \tau(T)$
2) $\tau(S \cap T) \subset \tau_i(\tau(S) \cap \tau(T))$.

Proof:

1) Since $S \cup T \subset \tau(S) \cup \tau(T)$, $\tau(S) \cup \tau(T) \subset \tau(S \cup T)$ and propagation in any layer network G_i cannot procede beyond $\tau(S) \cup \tau(T)$ by Lemma 3, we have $\tau(S \cup T) = \tau(S) \cup \tau(T)$.
2) Clearly $\tau(S \cap T) \subset \tau(S)$ and similarly $\tau(S \cap T) \subset \tau(T)$, hence

$$\tau(S \cap T) \subset \tau(S) \cap \tau(T) \subset \tau_i(\tau(S) \cap \tau(T))$$

for any i. □

Lemma 5: If the propagation model σ_i on each component of $\mathscr{G}$ is deterministic and submodular, the (deterministic) propagation σ in multiplex $\mathscr{G}$ will be submodular.

Proof: Let $i \in \{1, \ldots, n\}$. Then, by Lemmas 2, 3 and 4 and submodularity of σ_i

$$\begin{aligned}\sigma(S \cap T) + \sigma(S \cup T) &= |\tau(S \cap T)| + |\tau(S \cup T)| \\ &\leq |\tau_i(\tau(S) \cap \tau(T))| + |\tau_i(\tau(S) \cup \tau(T))| \\ &= \sigma_i(\tau(S) \cap \tau(T)) + \sigma_i(\tau(S) \cup \tau(T)) \\ &\leq \sigma_i(\tau(S)) + \sigma_i(\tau(T)) \\ &= \sigma(S) + \sigma(T).\end{aligned}$$

Thus, σ is submodular. □

b) Probabilistic case: Now the result for the deterministic case is generalized to the case when all networks satisfy GDS.

Theorem 1: Given multiplex network $\mathscr{G}$ with k layer networks G_i, if the model σ_i on each layer network G_i satisfies GDS then σ satisfies GDS.

Proof: Since σ_i satisfies GDS, with probability p_{ij}, G_i has deterministic, submodular propagation σ_{ij}, such that $\sigma_i = \sum p_{ij}\sigma_{ij}$. The probability that σ will comprise $\sigma_{1j_1}, \sigma_{2j_2}, \ldots, \sigma_{kj_k}$ is $\prod_{i=1}^{k} p_{ij_i}$, since propagation in each graph is independent. Let this propagation in $\mathscr{G}$ be labeled $\sigma_{j_1,\ldots,j_k}$. By Lemma 5, $\sigma_{j_1,\ldots,j_k}$ is submodular and deterministic. □

2) (1-1/e)-Approximation Algorithm: In this section, we detail the greedy algorithm ISF for solving the MIM problem. As shown in Algorithm 1, ISF is a greedy algorithm (with CELF++ optimization [15]), which chooses a node that maximizes the marginal gain of σ at each iteration. Recall that σ is the expected activation of the influence model σ defined on the multiplex in Section II, which incorporates the models σ_i on each layer utilizing the overlapping users. To compute σ, it is necessary to compute expected activation σ_i on each layer. To perform this computation, we use independent Monte Carlo simulations–in general, σ_i could be any model of influence propagation, and thus may not be amenable to specialized techniques for triggering model [7].

Algorithm 1 ISF: An Algorithm for Finding the Best Seed Users. Approximation Ratio: $1 - 1/e$ When Each Layer Satisfies GDS Property

```
Input: A multiplex $\mathscr{G} = (G^1, G^2, \ldots, G^k)$, $l$
Output: Seed set $S$ of size $l$
Renumber all the nodes across all networks so that each
    node gets a unique id
$S \leftarrow \emptyset$
$V \leftarrow \cup_{i=1}^{k} V_i$
for each $v \in V$ do
    $v.marginal\_gain = \sigma(v)$
    $v.round = 0$
end for
Initialize max priority queue $Q$ with (key,value) pair
    $(v, v.marginal\_gain)$, $\forall v \in V$
Initialize previous marginal gain, $prev\_mg = 0$
while $|S| \leq l$ do
    $v \leftarrow Q.pop().key$
    if $v.round == S.size$ then
        $S \leftarrow S \cup \{v\}$
        $prev\_mg \leftarrow prev\_mg + v.marginal\_gain$
    else
        $v.marginal\_gain \leftarrow \sigma(S \cup \{v\}) - prev\_mg$
        $v.round = S.size$
        $Q.add(v, v.marginal\_gain)$
    end if
end while
Return $S$
```

As shown earlier, σ is submodular and monotone increasing when each individual network satisfies GDS; therefore, in this case ISF has an approximation ratio of $(1 - 1/e)$ [12]. The time complexity of ISF is $O(nl(m + n)\log n)$ where n, m are number of users, friendships in the multiplex of OSNs, respectively. Each Monte Carlo simulation takes time $\Omega(n + m)$, and the $\log n$ factor accounts for the time to adjust the priority queue, l is the size of the seed set chosen.

B. Parallelizable Multiplex Algorithm

Although in the case that the model of propagation on each layer satisfies GDS, we have the $(1 - 1/e)$ performance guarantee of the greedy ISF, the running time of ISF may be impractical for large network sizes; hence, we propose Algorithm 2 (KSN), another approximation algorithm which parallelizes the problem in terms of the component layers–the difficulty lies in combining the solutions to the influence maximization problem on the separate layers to obtain a solution for MIM. KSN achieves this by approximating the solution to multiple-choice knapsack problem. The approximation ratio of KSN depends on the number of overlapping users o, the number of layers k, an arbitrary $\epsilon > 0$, and the ratio α of its input homogeneous layer algorithm A.

1) Description of KSN: The KSN algorithm takes as input an algorithm A (with ratio α) to solve the influence maximization problem on a single layer network, a multiplex network $\mathscr{G}$ with k layers, and number of seeds to pick l. For each $j \in \{1, \ldots, l\}$, $i \in \{1, \ldots, k\}$, algorithm A is run in parallel on each G_i to get seed sets T_{ij} with j seed nodes. It then uses an approximation to the multiple-choice knapsack problem (defined below) to decide how many nodes should be seeded in each layer, i.e., for each i, which T_{ij} to pick.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 2 KSN: A Knapsack Approach to Finding the Best Seed Users. Approximation Ratio: $((1-\epsilon)\alpha)/((o+1)k)$, Where $\epsilon > 0$, o is the Number of Overlapping Users, k is the Number of Layers, and α is the Ratio of Algorithm A on Homogeneous Networks

```
Input: Algorithm $A$, a multiplex network $\mathscr{G} = (G^1, G^2, \ldots, G^k)$, $l$
Output: Seed set $T$ of size $l$
for $i \in \{1, \ldots k\}$ do
    Run algorithm $A$ on $G_i$ with input $j$ to get seed sets $T_{i1}, T_{i2}, \ldots, T_{il} \subset G_i$, with $|T_{ij}| = j$.
end for
For each $T_{ij}$, let cost $c(T_{ij}) = |T_{ij}|$, and profit $p(T_{ij}) = \sigma(T_{ij})$.
Use $(1-\epsilon)$-approximation to multiple-choice knapsack problem (MCKP) to choose for all $i$, $T'_i \in \{T_{i1}, \ldots, T_{il}\}$, which choice satisfies $\sum_{i=1}^{k} |T'_i| = l$.
Return $T = \bigcup_i T'_i$.
```

-	0	1	2	3	4
G_1	0	200	350	400	425
G_2	0	600	601	602	603
G_3	0	200	210	214	214

Fig. 5. Example of how KSN works, as described in the text.

For example, suppose we have a multiplex with three layers: G_1, G_2, and G_3. Using algorithm A, we generate the table in Fig. 5, where the (i, j)th entry gives the activation of seeding j nodes in layer G_i. We then use an algorithm for multiple-choice knapsack to choose for each layer G_i, the number j_i of nodes to seed in that layer.

a) Worst case Bound on Performance of KSN: First, we need the definition of the MCKP problem.

Definition 5 (Multiple-choice knapsack problem): Let $(\mathscr{C}, k, l, c, p, B)$ be given, where $\mathscr{C} = \{C_1, \ldots C_k\}$ comprises k classes of l objects, $C_i = \{x_{ij} : 1 \leq j \leq l\}$, c and p are cost and profit functions on objects x_{ij}, and budget $B \geq 0$. The MCKP is to pick one item from each class, x'_i such that profit $\sum_{i=1}^{k} p(x'_i)$ is maximized under the constraint $\sum_{i=1}^{k} c(x'_i) < B$.

For $\epsilon > 0$, MCKP has a $(1-\epsilon)$-approximation as shown in [16]. We will use this algorithm to obtain an approximation for MIM as follows. Let an instance $(\mathscr{G}, k, l)$ of MIM be given. For each pair (i, j), $1 \leq i \leq k, 1 \leq j \leq l$, let T_{ij}^{opt} be an optimal seed set for G_i satisfying the two conditions $T_{ij}^{\text{opt}} \subset G_i$ and $|T_{ij}^{\text{opt}}| = j$. In addition, let T_{ij} be the approximation from algorithm A to T_{ij}^{opt}, that is, $T_{ij} \subset G_i$, $|T_{ij}| = j$, and

$$\sigma(T_{ij}^{\text{opt}}) \leq \alpha^{-1}\sigma(T_{ij}).$$

Then, let $C_i = \{T_{i0}, \ldots, T_{il}\}$, $C_i^{\text{opt}} = \{T_{i0}^{\text{opt}}, \ldots, T_{il}^{\text{opt}}\}$. Finally, let $\mathscr{C} = \{C_i : 1 \leq i \leq k\}$, $\mathscr{C}^{\text{opt}} = \{C_i^{\text{opt}} : 1 \leq i \leq k\}$, and for each i, j, define $c(T_{ij}) = j$, $p(T_{ij}) = \sigma(T_{ij})$, and likewise define c, p^{opt} for each T_{ij}^{opt}. Thus, we have two instances of the knapsack problem, namely, $I_1 = (\mathscr{C}, k, l, c, p, l)$ and $I_2 = (\mathscr{C}^{\text{opt}}, k, l, c, p^{\text{opt}}, l)$.

Lemma 6: Let Opt_{I_i} be the value of the optimal solution to MCKP instance I_i, $i \in \{1, 2\}$. Then,

$$\alpha\text{Opt}_{I_2} \leq \text{Opt}_{I_1}.$$

Proof: Suppose $\{T_{ib_i} : 1 \leq i \leq k\}$, $\{T_{ia_i}^{\text{opt}} : 1 \leq i \leq k\}$ are the optimal solutions for I_1, I_2, respectively. Then,

$$\begin{aligned}\alpha\text{Opt}_{I_2} &= \alpha \sum_i \sigma\left(T_{ia_i}^{\text{opt}}\right)\\ &\leq \sum_i \sigma(T_{ia_i})\\ &\leq \sum_i \sigma(T_{ib_i})\\ &= \text{Opt}_{I_1}\end{aligned}$$

since algorithm A's selection of T_{ia_i} ensures $\sigma(T_{ia_i}) \geq \alpha\sigma(T_{ia_i}^{\text{opt}})$. The last inequality follows from the fact that $\{T_{ia_i}\}$ is a feasible solution to instance I_1, and $\{T_{ib_i}\}$ is the optimal solution to I_1. □

Theorem 2: Let A be an α-approximation to the problem of influence maximization on a homogeneous single layer, and let o, k be the number of overlapping users and layers, respectively, in the multiplex. Furthermore, suppose the propagation σ_i on each layer of the multiplex is submodular. Then, KSN has approximation ratio $\frac{(1-\epsilon)\alpha}{(o+1)k}$.

Proof: Suppose KSN returns the union of $T_{1a_1}, T_{1a_2}, \ldots, T_{1a_k}$, selected from I_1. Let S_{opt} be the optimal solution to MIM instance $(\mathscr{G}, k, l)$. Let $\sigma(S_{\text{opt}})^i$ denote the expected activation under σ in layer G_i. Immediately, we have

$$\sigma(S_{\text{opt}}) \leq \sum_{i=1}^{k} \sigma(S_{\text{opt}})^i. \tag{1}$$

Also, letting O be the set of overlapping users, we have

$$\sigma(S_{\text{opt}})^i \leq \sigma_i(S_{\text{opt}} \cup O) \leq \sigma_i(S_{\text{opt}}) + \sigma_i(O) \tag{2}$$

where the first inequality in (2) follows from the fact that any activation in G_i proceeds according to the model σ_i and results from seed nodes in $S_{\text{opt}} \cap G_i$ or through overlapping users O. The second inequality in (2) follows from submodularity of σ_i.

Recall that OPT_{I_j} denotes the optimal value of MCKP on instance I_j, for $j = 1, 2$ as defined above; let KSN denote the value of the solution returned by Algorithm KSN. Then, $\frac{1}{1-\epsilon} KSN \geq \text{OPT}_{I_1}$; finally, notice if S is any set of size at most l, and i is a fixed layer

$$\frac{1}{(1-\epsilon)\alpha} KSN \geq \text{OPT}_{I_2} \geq \sigma_i(S) \tag{3}$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

by Lemma 6, and since $\sigma_i(S)$ is the value of a feasible solution to MCKP instance I_2. Therefore, by (1)–(3), we have

$$\begin{aligned}\sigma(S_{\text{opt}}) &\le \sum_{i=1}^{k}\sigma(S_{\text{opt}})^i\\ &\le \sum_{i=1}^{k}\sigma_i(S_{\text{opt}}) + \sum_{i=1}^{k}\sigma_i(O)\\ &\le \frac{k}{(1-\epsilon)\alpha}KSN + \sum_{i=1}^{k}\sigma_i(O)\\ &\le \frac{k}{(1-\epsilon)\alpha}KSN + \sum_{v\in O}\sum_{i=1}^{k}\sigma_i(v)\\ &\le \frac{k}{(1-\epsilon)\alpha}KSN + \frac{ok}{(1-\epsilon)\alpha}KSN\\ &\le \frac{(o+1)k}{(1-\epsilon)\alpha}KSN.\end{aligned}$$

□

b) Time complexity of KSN: KSN runs algorithm A in parallel $l \cdot k$ times, then employs the $(1-\epsilon)$ MCKP algorithm from [16]. Thus, if $tc(A, G_i, j)$ is the time complexity of A on j seed nodes with graph G_i, the time complexity of KSN is

$$O\left(\max_{(i,j)=(1,1)}^{(k,l)} tc(A, G_i, j) + (kl)^{\lceil 1/\epsilon - 1\rceil}\log k\right)$$

since $O((kl)^{\lceil 1/\epsilon-1\rceil}\log k)$ is the time complexity for the $(1-\epsilon)$ MCKP algorithm with k classes and l items in each class.

Notice that the scalability of KSN depends on the scalability of the input algorithm A. For example, in the special case that each σ_i satisfies the triggering model [11] and also satisfies each σ_i submodular, then letting algorithm A be the TIM algorithm from [7], we would have the expected running time of KSN bounded by

$$O((k+\ell)(m+n)\log n/\epsilon^2 + (kl)^{\lceil 1/\epsilon-1\rceil}\log k)$$

where n is the maximum number of nodes in a layer, m is maximum number of edges in a layer, ℓ is an integer; and approximation ratio

$$\frac{(1-\epsilon)(1-1/e-\epsilon)}{(o+1)k}$$

with probability $(1-n^{-\ell})^k$.

IV. Experimental Results

In this section, we perform experiments on both synthesized and real-world networks to show the effectiveness of the proposed algorithms.

A. Methodology

We evaluated the following algorithms:

1) ISF (Algorithm 1), the greedy algorithm with CELF++ optimization on the multiplex;
2) KSN (Algorithm 2), with algorithm A in the definition of Algorithm 2 is set to the CELF++ algorithm [15] or the IMM algorithm [7];
3) even seed (ES), which seeds each layer of the multiplex with an equal number of seed nodes l/k;
4) best single network (BSN), which places all l seed nodes in the layer that maximizes $\sigma_i(S_i)$, where S_i is the seed set chosen according to CELF++ in layer i, with $|S_i| = l$.

To estimate the expected activation σ on the multiplex, or σ_i in layer G_i, we use independent Monte Carlo simulations.

Since the greedy CELF++ approach is not very scalable, we limit the maximum length of a diffusion sequence to 4 in Sections IV-B and IV-C. The experiments in Sections IV-B and IV-C were run on a machine with an Intel(R) Xeon(R) W350 CPU and 12-GB RAM.

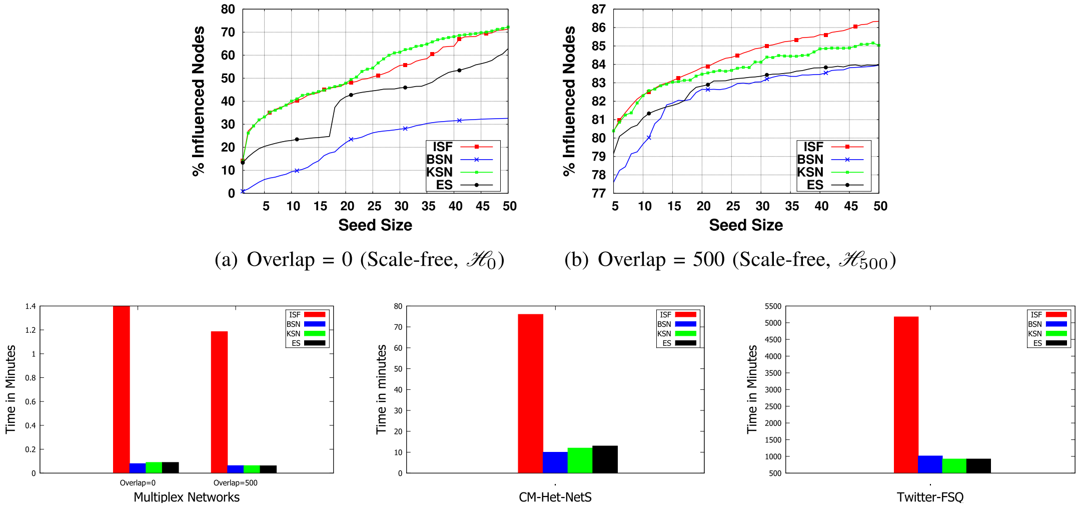
In Section IV-D, we demonstrate the scalability of KSN on large multiplexes. This implementation[1] of KSN is parallelized and utilizes algorithm A set to the IMM algorithm of Tang [18]. We chose IMM since it is highly scalable and source code is available to solve the single layer problem with both IC and LT models. For the MCKP problem within KSN, we used our implementation of the 1/2-approximation from Chandra *et al.* [16]. These experiments were run on a machine with 2 Intel(R) Xeon(R) CPU E5-2697 v4 2.30-GHz and 384-GB RAM.

B. Synthesized Multiplexes

We consider synthesized multiplexes based on three scale-free (SF) networks H_1, H_2, and H_3 generated according to Barabasi–Albert model [18] with 1000 nodes and 4000 edges, with average degree 4; the exponent in the power-law degree distribution generated by this method is 2, which is consistent with that observed in real-world social networks which have exponents 2–3, hence, these synthesized networks should act as a good representative for capturing the social influence spread phenomenon. We assigned H_1, H_2, and H_3 with diffusion models LT, IC, and majority linear threshold (MLT), respectively, with edge weights and thresholds chosen uniformly in [0, 1], where MLT is the deterministic, nonsubmodular majority linear threshold model [19], whereby a node is activated if majority of its neighbors are activated; then, to form the multiplexes, beginning with a specified number of overlapping users o, we select the overlapping users randomly, such that each overlapping user is in all three of the layers, that is, to create an overlapping node, we randomly choose indices from each layer, and add two interlayer edges to connect these three separate users into a single overlapping user. This step is repeated until we have o overlapping users in the multiplex. A multiplex created in this way will be called a SF multiplex, and will be denoted $\mathscr{H}_o$, where o is the number of overlapping users.

1) Algorithm Performance on Synthesized Multiplex Networks: The performance of all four algorithms on the SF multiplex $\mathscr{H}_0$ with no overlapping users is shown in Fig. 6(a). As may be suggested by the analysis of the performance ratio for KSN, which depends on the number of overlapping users $o = 0$, the performance of KSN equals or exceeds ISF.

[1] Source code available at http://www.alankuhnle.com/papers/mim/mim.html



(a) Overlap = 0 (Scale-free, $\mathscr{H}_0$)

(b) Overlap = 500 (Scale-free, $\mathscr{H}_{500}$)

(c) Scale-free, $\mathscr{H}$

(d) CM-Het-NetS

(e) Twitter-FSQ

Fig. 6. (a) and (b) Comparison of the four methods with no overlap and 500 overlapping users, respectively, in the SF multiplex setting. (c)–(e) Running time comparison for the four algorithms on different multiplexes.

ES, where seeds are split evenly among the three layers, requires 40 seed users to get comparable activation to KSN and ISF at 20 seed users. BSN does the worst since it is choosing seed users from a single layer– since all three layers have 1000 nodes, BSN cannot activate more than 33% of the multiplex, which it approaches as the number of seed nodes $l \geq 35$.

Next, we considered the $\mathscr{H}_{500}$, the multiplex with the same layers but 500 overlapping users, which is 1/6 of the original 3000 users. The performance is shown in Fig. 6(b). In the case of this significant overlap, ISF outperforms KSN. BSN is no longer limited to activation of at most 33% and performs similar to ES.

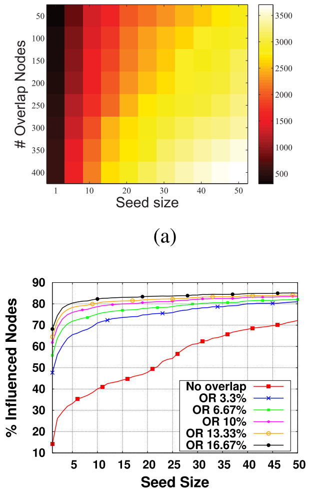
These results on the synthesized SF multiplex demonstrate how, in the case of small overlap we expect KSN to perform as well or better than ISF; however, as overlap increases, the performance of KSN will degrade with respect to ISF, a statement that we have demonstrated theoretically in Section III-B1.a

2) Running Time: In Fig. 6(c), we compare the running time for the four algorithms on $\mathscr{H}_0$ and $\mathscr{H}_{500}$. The effect of parallelization by layer of the multiplex on the running time may be seen; note that for KSN, BSN, and ES, we are using CELF++ on each layer. This algorithm could be replaced by a more efficient single layer algorithm, which would further improve running time as compared to ISF.

3) Role of Overlapping Users:

a) On total activation: To investigate the role of overlapping users further, we varied the number of overlapping users from 50 to 400 in the ER multiplex setting. The effect of this on the total activation of ISF can be seen in Fig. 7(a). Increasing the number of overlapping users increases the number of influenced users.

b) On performance of KSN: We have already noticed, both from theoretical and experimental standpoints, that the performance of KSN with respect to the optimal solution is expected to degrade as the number of overlapping users increases. In this section, we examine how the performance of KSN compares with itself on the SF synthesized multiplexes as the number of overlapping users increases. The results are shown in Fig. 7(b). As the number of overlapping users increases, the algorithm's performance improves drastically, demonstrating the efficacy of KSN even when overlapping users exist. With the overlapping percentage at 16.67%, KSN activates over 80% of the nodes in the multiplex with just five seed nodes, as opposed to in the case of no overlap, where activation is at roughly 35% with five seed nodes. In addition, this experiment provides further evidence of the strong benefit overlapping users provide in the influence propagation.



(a)

(b)

Fig. 7. Influence Spread for different overlap ratio, diffusion models– Net1: IC, Net2: LT. (a) Total activation of ISF: overlap size versus seed size (Erdos-Renyi). (b) Effect of overlapping on KSN performance (Scale free).

C. Algorithm Performance on Real Networks

The first real multiplex we consider is based upon sections of *Twitter* and *Foursquare* (*FSQ*) networks. The generation of this data set is described in [20]; overlapping users were

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

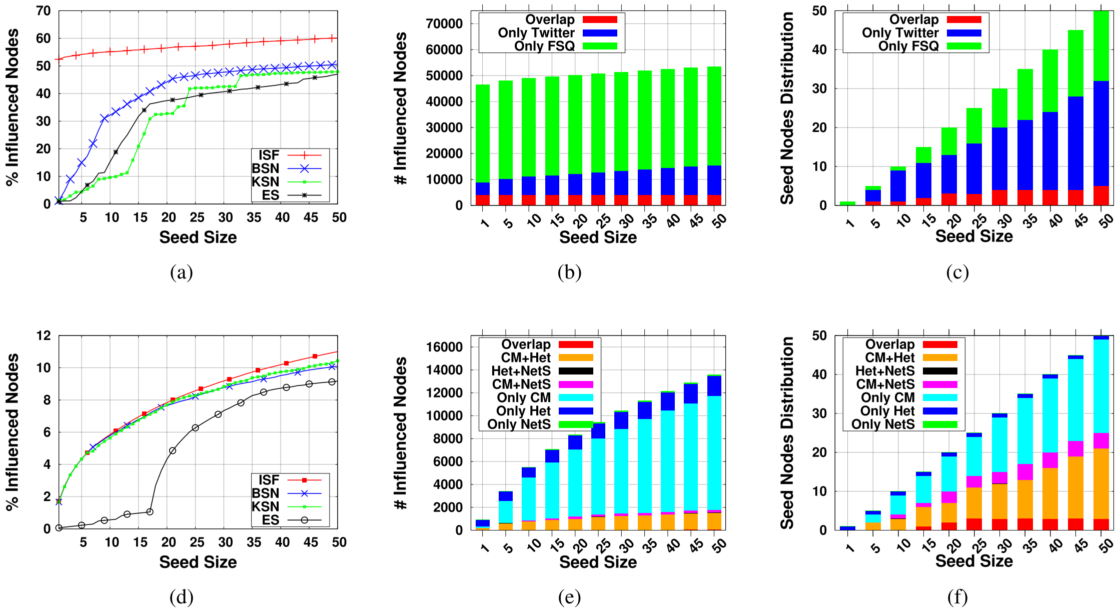


Fig. 8. Twitter-FSQ Network (top), co-author networks (bottom). (a) Total activation in Twitter-FSQ. (b) Activated node composition (Twitter-FSQ). (c) Seed node composition (Twitter-FSQ). (d) Total activataion in CM-Het-NetS. (e) Activated node composition (CHN). (f) Seed node composition (CHN).

TABLE I
TRACES OF REAL NETWORKS

Networks	Nodes	Edges	Avg Deg
Twitter	48277	16304712	289.7
FSQ	44992	1664402	35.99
CM	40420	175692	8.69
Het	8360	15751	1.88
NetS	1588	2742	1.73

identified by using Foursquare API v1 to identify the Twitter usernames corresponding to a Foursquare account. The weight of each link in Twitter is inferred by using frequency of tweets between users. In Foursquare, the weight of each link is assigned value 1, due to the lack of a message data set. The number of overlapping nodes is 4100, see Table I.

The second real multiplex is based upon academic collaboration networks, described in [20]. The layers are organized by the research area: condensed matter (CM), high-energy theory (Het) [21], and network science (NetS) [22]. A user is considered to be overlapping if he or she has published in two or more of these three fields. The CM-HET-NetS (CHN) multiplex is considered an undirected network throughout the experiments. The number of overlapping users are 2860, 517, and 90 between CM-HET, CM-NetS, and HET-NetS, respectively; 75 users are present in all three networks.

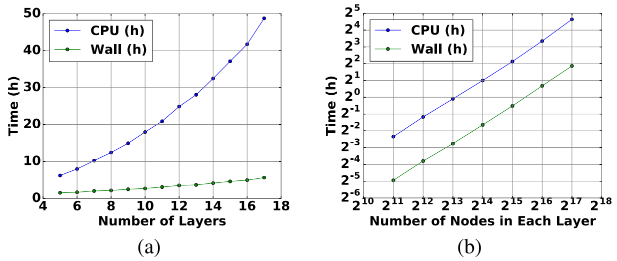
1) Model Selection: Saito *et al.* [2] have performed machine-learning techniques to match variants of the IC and LT models with real propagation events. They found that even on the same network, different propagation events may be better explained by disparate models. In all the Twitter-FSQ experiments we assigned Twitter the LT model, and Foursquare the IC model; experiments that swapped the model selection gave similar qualitative results. On CHN, we assigned CM, HET, and NetS for the LT, IC, and LT models, respectively. Thresholds were assigned uniformly randomly in [0, 1], while edge weights were determined as described above for Twitter, and randomly in [0, 1] for the other networks. Notice that with this choice of models, Theorem 1 applies; therefore, the approximation ratios of ISF and KSN hold.

It is evident from Fig. 8(a) that the seeds found by ISF in Twitter-FSQ network outperforms BSN (20% larger for $l = 50$) as well as ES and KSN. An interesting observation in this figure is that relatively few (overlapping) nodes are responsible for a lot of the propagation in the ISF case–the seed node composition is shown in Fig. 8(c). Also, in Fig. 8(b), we see influence spread is larger in FSQ compared to Twitter.

Since the condensed matter network is comparatively larger than the other two, most of the seed users are selected from it as shown in Fig. 8(f). Therefore, a significant number of finally activated users also reside in this network. The influence spread obtained in the multiplex network only taking the seeds of BSN and KSN are very close to that obtained by the seed nodes identified by ISF. Nonetheless, ISF outperforms them as the seed set becomes larger, as shown in Fig. 8(d); this again illustrates the benefit of taking into account overlapping users in the solution.

As depicted in Fig. 8(b), the composition of influenced users in the Twitter-FSQ network suggests that majority of influenced users in the multiplex network belongs to FSQ which implies propagation spreads easily in this network compared to Twitter. The same observation can be drawn for CM network in case of co-author network, shown in Fig. 8(e). As illustrated in Fig. 8(c) and (f), the seed set of the multiplex network identified by ISF contains a much higher number of nodes from a specific network than other networks. In addition,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a) (b)

Fig. 9. Scalability of KSN. (a) $n_{\mathrm{ER}} = 10^5$. (b) $k = 10$.

overlapping users show significant role in diffusing information by occupying a considerable fraction of the seed set chosen by ISF.

a) Running time: In Fig. 6, we compare the running time for the four algorithms on CHN, and Twitter-FSQ multiplexes. The effect of parallelization by layer on the running time may be seen, with ISF taking much longer in all cases.

D. Scalability of KSN

In our final set of experiments, we demonstrate the scalability of KSN on large multiplexes. For these experiments, we used synthesized multiplexes, where each layer is an ER network with average degree 5. Each layer has the same number n_{ER} of vertices. Layer i is assigned model IC if i is even and model LT otherwise, with edge weights uniformly chosen in $(0, 0.1)$. The number of overlapping users is set to $o = 0.1\ n_{\mathrm{ER}}$, and each overlapping user is present in all layers of the multiplex. For these experiments, the number of seeds $l = 100$. Since the solution of IMM of size m is not necessarily contained in the solution of size $m + 1$, IMM is run on each layer for all values from 1 to l, as indicated in the pseudocode of KSN.

Results for the running time of KSN are shown in Fig. 9. The first experiment varies the number k of layers in the multiplex from 5 to 17, with $n_{\mathrm{ER}} = 10^5$. Thus, the largest multiplex in this experiment has 1.54×10^6 unique users. In Fig. 9(a), we show the running time in hours of KSN versus k, for both total CPU time and the wall-clock time. Thus, on the largest multiplex with $k = 17$, KSN finished in roughly 5 h of wall-clock time, demonstrating a high level of parallelization. Results for the second experiment are shown in Fig. 9(b), where the number of layers is fixed at $k = 10$, and n_{ER} is varied from 2048 to 131072. In this experiment, both the wall-clock time and CPU time increase linearly, with the wall-clock time below 4 h on the largest multiplex with 1.19×10^6 unique users.

V. Related Works

In a seminal work on single-layer networks, Kempe *et al.* [11] showed the IC and LT models were submodular by utilizing a "live edge" approach, thereby allowing the use of a greedy algorithm to approximate the influence maximization on single networks. The "live edge" approach implicitly establishes the stronger GDS property as defined above for these models. The inapproximability of the classical max coverage problem precludes any better approximation to the influence maximization problem. Since this work, there have been a number of improvements to the running time of the greedy algorithm on single networks. The first improvement was through the use of priority queue to do a "lazy evaluation" of estimated influence in [23]. Cohen *et al.* [5] provided an approximation algorithm for influence maximization one or two orders of magnitude faster than previous works on single-layer Networks; Kuhnle *et al.* [25] adapted this framework for the threshold activation problem. Borgs *et al.* [6] provide a nearly runtime-optimal algorithm on single networks with the IC model, and Tang *et al.* [7] provided a fast algorithm with with high probability achieves the greedy ratio $1 - 1/e - \epsilon$ for the triggering model; this method was further improved by Tang [17] using martingales. Nguyen *et al.* [25], [26] and Huang *et al.* [27] have further improved the sampling techniques to yield even faster algorithms for the single-layer network problem with the $1 - 1/e - \epsilon$ ratio; also, Li *et al.* [28] have developed a scalable and nearly optimal algorithm.

A considerable number of works have studied influence maximization for variants of IC models and its extensions such as [3] and [4]. For a deterministic variant of LT, Zou *et al.* [29] showed NP-completeness for the problem and Dinh *et al.* [30] proved the inapproximability as well as proposed efficient algorithms for this problem on a special case of LT model. In their model, the influence between users is uniform and a user is influenced if a certain fraction ρ of his friends are active.

Researchers have started to explore the influence maximization problem on multiplex networks with works of Yagan *et al.* [31] and Liu *et al.* [32], who studied the connection between off-line and online networks. The first work investigated the outbreak of information using the SIR model on random networks. The second one analyzed networks formed by online interactions and off-line events. The authors focused on understanding the flow of information and network clustering but not solving the heterogeneous influence problem.

Shen *et al.* [20] explored the information propagation in multiplex OSNs taking into account the interest and engagement of users. The authors combined all networks into one network by representing an overlapping user as a super node. This method cannot preserve the heterogeneity of the layers. Nguyen *et al.* [33] studied the influence maximization problem, which handles multiple networks but only considers homogeneous diffusion process across all the networks. None of these works took into consideration the heterogeneity of diffusion processes in multiplex networks. On the other hand, our scheme overcomes these shortcomings by enabling different networks to have different influence propagation models. Pan *et al.* [34] studied threshold activation problems on multiplex networks under a diffusion model with continuous time.

VI. Conclusion

We formulate the MIM problem that seeks to maximize influence propagation in a multiplex with overlapping users and heterogeneous propagation. We provide a property (GDS),

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

which carries over from single layer to multiplex propagation, giving the $1 - 1/e$ ratio for the greedy algorithm ISF if it is satisfied on each layer. We also develope an approximation algorithm KSN that benefits from the optimizations that influence maximization in a single network has undergone (e.g., [5]–[7]). We prove the approximation ratio of KSN, which depends on the number of overlapping users. As demonstrated in our experimental section, the performance KSN may fall short of ISF when overlapping effects become large. Due to the long running time of ISF, future work would include attempting to find faster approximation algorithms in the heterogeneous multiplex setting.

References

[1] *216 Social Media and Internet Statistics*. [Online]. Available: http://thesocialskinny.com/216-social-media-and-internet-statistics-september-2012/

[2] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Selecting information diffusion models over social networks for behavioral analysis," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), vol. 6323. Cham, Switzerland: Springer, 2010, pp. 180–195.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *Proc. ICALP*, 2005, pp. 1127–1138.

[4] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. KDD*, 2010, pp. 1029–1038.

[5] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 629–638.

[6] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. 25th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2014, pp. 946–957.

[7] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. 2014 ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2014, pp. 75–86.

[8] *Overlap Among Major Social Network Services*. [Online]. Available: http://www.tomhcanderson.com/2009/07/09/overlap-among-major-social-network-services/

[9] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *Proc. ICWSM*, 2011, pp. 522–525.

[10] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Discovering links among social networks," in *ECML PKDD*. 2012, pp. 467–482.

[11] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. KDD*, 2003, pp. 137–146.

[12] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Programm.*, vol. 14, no. 1, pp. 265–294 1978.

[13] W. Chen, L. V. S. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," in *Synthesis Lectures on Data Management*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[14] H. Li, S. S. Bhowmick, A. Sun, and J. Cui, "Conformity-aware influence maximization in online social networks," *VLDB J.*, vol. 24, no. 1, pp. 117–141, 2015.

[15] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 47–48.

[16] A. K. Chandra, D. S. Hirschberg, and C. K. Wong, "Approximate algorithms for some generalized knapsack problems," *Theor. Comput. Sci.*, vol. 3, no. 3, pp. 293–304, 1976.

[17] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1539–1554.

[18] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[19] N. Chen, "On the approximability of influence in social networks," in *Proc. SODA*, 2008, pp. 1029–1037.

[20] Y. Shen, T. N. Dinh, H. Zhang, and M. T. Thai, "Interest-matching information propagation in multiple online social networks," in *Proc. CIKM*, 2012, pp. 1824–1828.

[21] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.

[22] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 3, p. 036104, 2006.

[23] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. KDD*, 2007, pp. 420–429.

[24] A. Kuhnle, T. Pan, M. A. Alim, and M. T. Thai, "Scalable bicriteria algorithms for the threshold activation problem in online social networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, May 2017, pp. 1–9.

[25] H. T. Nguyen, T. N. Dinh, and M. T. Thai, "Cost-aware targeted viral marketing in billion-scale networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[26] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proc. ACM SIGMOD/POSD Conf.*, 2016, pp. 695–710.

[27] K. Huang, S. Wang, G. Bevilacqua, X. Xiao, and L. V. S. Lakshmanan, "Revisiting the stop-and-stare algorithms for influence maximization," *Proc. VLDB Endowment*, vol. 10, no. 9, pp. 913–924, May 2017.

[28] X. Li, J. D. Smith, T. N. Dinh, and M. T. Thai, "Why approximate when you can get the exact? Optimal targeted viral marketing at scale," in *Proc. IEEE Int. Conf. Comput. Commun.*, May 2017, pp. 1–9.

[29] F. Zou, Z. Zhang, and W. Wu, "Latency-bounded minimum influential node selection in social networks," in *Proc. WASA*, 2009, pp. 519–526.

[30] T. N. Dinh, D. T. Nguyen, and M. T. Thai, "Cheap, easy, and massively effective viral marketing in social networks: Truth or fiction?" in *Proc. 23rd ACM Conf. Hypertext Social Media (HT)*, 2012, pp. 165–174.

[31] O. Yagan, D. Qian, J. Zhang, and D. Cochran, "Information diffusion in overlaying social-physical networks," in *Proc. CISS*, Mar. 2012, pp. 1–6.

[32] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: linking the online and offline social worlds," in *Proc. KDD*, 2012, pp. 1032–1040.

[33] D. T. Nguyen, H. Zhang, S. Das, M. T. Thai, and T. N. Dinh, "Least cost influence in multiplex social networks: Model representation and analysis," in *Proc. ICDM*, Dec. 2013, pp. 567–576.

[34] T. Pan, A. Kuhnle, X. Li, and M. T. Thai, "Popular topics spread faster: New dimension for influence propagation in online social networks," in *Proc. Int. Conf. Data Mining (ICDM)*, 2017, pp. 1–11.

Alan Kuhnle (S'17) received the M.Sc. degree in mathematics from the University of Florida, Gainesville, FL, USA, in 2013, where he is currently pursuing the Ph.D. degree in computer science.

His current research interests include approximation algorithms and combinatorial optimization, especially for complex networks, algorithms that can run on billion-scale networks, such as large biological or social networks, and dynamic networks, uncertain networks, online algorithms, and adaptive algorithms.

Mr. Kuhnle was a recipient of the University of Florida Graduate School Fellowship Award and the Post-Doctoral Fellowship Award from the University of Florida Informatics Institute.

Md Abdul Alim received the B.Sc. degree in computer science and engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, and the Ph.D. degree in computer science from the University of Florida, Gainesville, FL, USA, in 2016.

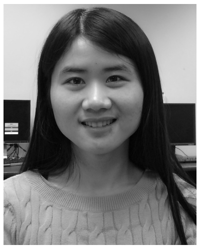
His current research interests include network vulnerability and community structure analysis in complex networks, including large-scale biological, wireless, and online social networks, social-aware device-to-device communication, influence propagation in online social networks, and approximation algorithms and its application in combinatorial optimization.

Dr. Alim was a recipient of several awards including the University of Florida Graduate School Fellowship Award, the Gartner Group Info Tech Fund, the NSF Travel Fellowship Award, and the CISE Student Travel Award.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Xiang Li received the M.Sc. degree from the Academy of Mathematics Systems and Science, Chinese Academy of Sciences, Beijing, China, in 2012, and the M.Sc. degree in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, where she is currently pursuing the Ph.D. degree with the Department of Computer and Information Science and Engineering.

Her current research interests include online social networks, network vulnerability, algorithms, and security in smart grid.



Huiling Zhang received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China. She is currently pursuing the Ph.D. degree at the University of Florida, Gainesville, FL, USA.

Her current research interests include approximation algorithms and its application in complex networks, particularly misinformation detection, network vulnerability, and influence propagation in online social networks.

My T. Thai (M'06) is currently a University of Florida Research Foundation Professor with the Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL, USA. Her studies have led to six books and over 140 articles published in leading journals and conferences. She has been involved in many professional activities. Her current research interests include scalable algorithms, big data analysis, cybersecurity, and optimization in network science and engineering, including communication networks, smart grids, social networks, and their interdependency.

Dr. Thai was a recipient of many research awards including the UF Provosts Excellence Award for Assistant Professors, the University of Florida Research Foundation Professorship Award, the Department of Defense Young Investigator Award, and the National Science Foundation CAREER Award. She received the IEEE MSN 2014 Best Paper Award and the 2017 IEEE ICDM Best Papers Award, for her research. She has been the TPC Chair for many IEEE conferences. She served as an Associate Editor for the *Journal of Combinatorial Optimization*, the *Journal of Discrete Mathematics*, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and the IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, and a Series Editor of the *Springer Briefs in Optimization*. She has co-founded and is a co-Editor-in-Chief of the *Computational Social Networks* journal.