Disrupting Diffusion: Critical Nodes in Network*

Raj Gaurav Kumar
Dept. of Computer Science
Iowa State University
gaurav@iastate.edu

Preeti Bhardwaj
IBM
preetibh@iastate.edu

Samik Basu
Dept. of Computer Science
Iowa State University
sbasu@iastate.edu

A. Pavan
Dept. of Computer Science
Iowa State University
pavan@iastate.edu

***Abstract*—We formulate and study the problem of identifying nodes whose absence can maximally disrupt network-diffusion under the independent cascade model. We refer to such nodes as *critical nodes*. We present the notion of *impact* and characterize critical nodes based on this notion. Informally, impact of a set of nodes quantifies the necessity of the nodes in the diffusion process. We prove that the impact is monotonic. Interestingly, unlike similar formulation of critical edges in the context of Linear Threshold diffusion model, impact is neither submodular nor supermodular. Furthermore, we prove that the problem of finding a set of nodes which maximizes impact is NP-Hard. Hence, we develop heuristics that rely on submodular approximations of the impact function. We empirically evaluate our heuristics by comparing the level of disruption achieved by identifying and removing critical nodes as opposed to that achieved by removing the most influential nodes.**

***Index Terms*—Information Diffusion, Critical Nodes, Social Networks, Submodularity**

I. Introduction

Two of the widely studied problems in the context of spread/diffusion (of information/opinions/disease) in complex networks involve (a) *influence maximization problem*—finding the set S of entities, called *seed set*, such that when the information originates from S, its diffusion in the network is maximal [5], [14]; (b) *source identification problem*—once the diffusion has occurred, identify a set of entities that can be classified as source/seed of the diffusion [12], [18], [26]. Addressing influence maximization problem results in finding a seed set, called *max seed*, of entities that can cause maximal information spread. Whereas source identification leads to identifying a possible seed that caused the observed spread.

Motivating Problem. In this work, we study a problem that is orthogonal to both of the above problems: *identify a set of size k of entities, which when removed from the network, maximally disrupts the diffusion of influence that may have started at any seed set*. More formally, the goal is to identify a set of nodes C such that, after removal of C from the network, $\sigma(S)$ (expected number of nodes that are influenced by seed set S) is maximally reduced for all S. We refer such entities C as *critical nodes*, and the problem of computing such nodes in the context of probabilistic diffusion as the *identifying critical nodes* (ICN) problem. The importance of addressing this problem cannot be understated. In social networks, influence of un-founded opinions or propagation of fake news can be avoided by identifying and informing/isolating the critical nodes. In computer network security, protecting critical nodes from known worms (via patching, security updates) can help in protecting the critical network-infrastructure from repeated disruption due to worm-attacks. In the context of disease propagation, helping critical communities that were once impacted by epidemics can make a difference in overall health of the population.

This paper contains work done by Preeti Bhardwaj and Raj Gaurav Kumar, when they were graduate students at ISU. Research supported in part by NSF grants 1849053 and 1934884

Note that, the critical nodes are not necessarily the max-seed or the source of diffusion; rather the critical nodes can be viewed as the ones whose presence is "critical" for diffusion in the network. In other words, criticality of a nodes can be described equivalently as how their presence is important for maximizing diffusion or (conversely) how their absence is important for minimizing diffusion. To illustrate the unique nature of critical nodes, consider the example network in the Figure 1. Let the objective be to identify one critical node. Directed edges in the network indicate that the influence diffuses from the the source to the destination, and the edge annotations capture the probability of the diffusion. Such a diffusion model is referred to as the independent cascade (IC) model, which directly captures the notion that new information/behaviors are contagious [14], [16]. Following the IC model, each node gets one chance to influence its neighbors. Assume that $m >> n$, edge probabilities to/from v_5 are close to 0 and all other probabilities are close to 1. Now, the most influential node is v_0 as it can influence almost the entire network (except the ones $u_1 \dots u_m$). However, removing v_0 does not disrupt the influence diffusion if some other seed is chosen. For instance, any one of v_2, v_3, or v_4 can still act as a source of influence that spreads to the same extent. The critical node, in this network, is v_4; removal of v_4 will maximally disrupt information diffusion from any other node.

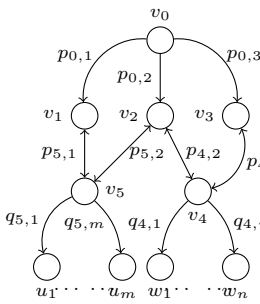


Fig. 1. Example

Our Solution. One of the primary challenges in addressing the ICN problem is that there may not be any solution (i.e., critical nodes), whose removal maximally reduces the diffusion originating from *all* seeds. Consider the ICN problem with $k = 1$: Let v_1 and v_2 be two different nodes such that removing v_1 maximally reduces the diffusion from a seed S_1 and removing v_2 maximally reduces the diffusion from another

seed S_2. Then there is no single node that maximally reduces the diffusion from S_1 and S_2.

We, therefore, characterize criticality by introducing the notion of *impact* of a set of nodes. Intuitively, impact of a set of nodes S quantifies the reduction in the expected diffusion from all nodes when the set S is removed from the network. That is, rather than reviewing the reduction in the expected diffusion from each seed set, we consider the reduction in the expected diffusion from all nodes. Consequently, higher impact of set of nodes implies higher criticality of the set. We formalize the ICN problem as finding a set of nodes with maximal impact.

We prove that impact is monotonic and is neither submodular nor supermodular. As a result, greedy algorithm applied to optimization of impact does not provide usual $(1 - 1/e)$ approximation guarantees as it does when applied to address different variations of influence maximization and source detection problems. We further prove that the problem of maximizing impact is NP-Hard. Given the hardness of the problem, greedy algorithm is still a viable strategy, where the impactful set is computed assuming submodularity of the impact function. Using the notion of random reachable sets from Borgs *et al.* [3], we obtain an efficient greedy algorithm, CRIT-SET, to compute high impact nodes.

We empirically validate that high impact nodes are indeed critical nodes as follows. First, we consider the problem of disrupting the influence from a given seed set. In this setting, we introduce variants of our heuristic, where criticality (or more precisely, impact) of nodes are computed in the context of the given seed, i.e., impact of node that disrupts diffusion from the seed is higher than the impact of nodes that do not disrupt diffusion from the seed. We identify the set k of high impact nodes and evaluate whether their removal indeed disrupts diffusion from the given seed. We compare our strategy against the baseline strategy, TOP-INFL, where the most influential nodes (other than the seed) are removed. We show that removal of high impact set of size k (as per CRIT-SET) causes more reduction in the influence than the removal of k influential nodes (as per TOP-INFL). We consider another variant TOP-CRIT of our heuristic, where the impact function is assumed to be modular; computing the set of nodes using TOP-CRIT is less expensive that computing using CRIT-SET. We show that while the results obtained using CRIT-SET and TOP-CRIT are both better than that obtained using TOP-INFL in the context of a given seed, the objective of reducing the influence from *any* seed is truly addressed by the set of critical nodes as computed using CRIT-SET.

II. RELATED WORK

Kempe et al. [14] proved that the problem of influence maximization in complex network is NP-hard and that greedy strategy has a $(1 - 1/e)$ approximation guarantee. The guarantee relies on diffusion being non-negative, monotonic and submodular. Several subsequent work focused on efficient implementation of the greedy strategy [4]–[10], [13], [19], [21], some of which do not admit to the same approximation guarantee. Recently, Borgs et al. [3] introduced an efficient technique based on random reachable sets to realize the greedy strategy with approximation guarantees. The technique was further refined by Tang et al. [23]. We will also use the strategy of random reachable sets to develop an efficient greedy heuristic to identify the critical nodes (Recall that we have already presented that influential nodes are not necessarily the critical nodes).

Khalil et al. [15] focus on removing edges for disrupting diffusion in linear threshold (LT) model. They prove that the function $f(E) = \sum_{v \in V} \sigma_{G/E}(v)$, where E is a set of edges and G/E corresponds to the network G with edges in E removed, is a supermodular function. This implies that, the critical edge identification problem in LT model reduces to maximizing a submodular function. In contrast, we will prove that if independent cascade diffusion model is considered, the optimization function is neither submodular nor supermodular. Another technique proposed in the context of LT model for disrupting diffusion involves using competitive diffusion. [11] focuses on competitive linear threshold model for diffusion disruption, where two types of diffusions propagate in the network; the objective is to identify a seed set for one type of diffusion (positive information) such that it minimizes the affect of the other type of diffusion (negative information).

Ventresca and Aleman [24] referred to critical nodes as the ones whose removal results in minimal pair-wise connectivity of the residual graph. They considered non-probabilistic network. As a result, minimizing pair-wise connectivity does not correspond to maximizing the disruption of probabilistic diffusion. Variants of critical node removal in the context of minimizing graph connectivity is also considered by Pullan [22].

The work done by Aspnes et al. [2] focus on identifying the nodes in the network, which when vaccinated, will contain the diffusion. Such nodes can be viewed as critical nodes in our setting. The authors present a game-theoretic formulation of the problem, develop a reduction to a graph partitioning problem and provide a poly-time greedy approximation algorithm. However, the authors assumed a simplistic diffusion model, where each active node deterministically activates its susceptible neighbors. This assumption along with the nature of the greedy strategy for partitioning does not make the process a feasible technique in the context of large social networks, where diffusion is probabilistic. Similarly, Kuhlman et al. [17] considered deterministic (non-probabilistic) diffusion model (threshold based model) for removing critical nodes in the context of random seed set with the objective to maximize the number of un-influenced node.

III. FORMALIZING CRITICALITY

Background. We present some of the basic definitions in the context of information diffusion in network. A network $G = (V, E)$, where V is a finite set of nodes and $E : V \times V \rightarrow [0, 1]$ is a directed edge relation between nodes annotated with a probability measure. The direction in the edge $u \xrightarrow{p_{u,v}} v$ indicates the direction of diffusion from u to v

and the annotation $p_{u,v}$ indicates the probability (*propagation probability*) of that diffusion. An undirected edge can be viewed as bi-directional with the same propagation probability in both directions. Each node in the network can be in two states: inactive (idle or susceptible) and active (influenced or infected); a node can evolve from being inactive to active and an active node remains active. In this work, we concentrate on Independent Cascade (IC) model, where at every (discrete) time step i, each node u, which is *newly activated* at time step $i-1$, will activate each of its (inactive) neighbor v (connected by a directed edge) with probability $p_{u,v}$. This captures diffusion at the i-th step. The diffusion process continues till no new node is activated.

Given a seed $S \subseteq V$, $\sigma(S)$ denotes the *expected* number of nodes influenced at the end of diffusion (we omit the subscript G, when the network information is immediate in the context). For example, $\sigma(v_0)$ in Figure 1 is $5+n$.

A. Critical Nodes as Impactful Nodes

We introduce the concept of *impact* of node(s) and claim that impact can be used effectively to compute the criticality of node(s). We first present the notion of *strength of diffusion*.

Definition 1. *Given a network $G=(V,E)$, the strength of diffusion in G, denoted by $\texttt{ST}(G)$, is* $\sum_{v\in V}\sigma_G(v)$.

For example, for the graph in Figure 1, if all probabilities except the ones to/from v_5 (which are close to 0) are equal to 1, then the strength of diffusion is $(5+n)+1+3\times(3+n)+n+1+m = 16+5n+m$. Intuitively, the strength of diffusion indicates sum of the expected number of nodes each node may influence. If the strength of diffusion in a network is high, then it indicates that the network has "many nodes" that can influence a lot of nodes of the network. This can be interpreted as: the network has many good seed sets that can collectively influence a large population of the network. Conversely, if the strength of influence is small, it is an indication that there are no (or very few) seed sets having high influence. If removal of a set of nodes from a network causes the strength of diffusion to go down, then it indicates the influence of all (or many) seed sets is also reduced. Thus a set of nodes whose removal will cause maximal reduction in the strength of diffusion can be considered as critical nodes. We introduce impact as the decrease in the strength of diffusion of network.

Definition 2 (Impact of Node(s))**.** *Given a network $G=(V,E)$, the impact of $S \subseteq V$, denoted by $\texttt{IM}_G(S)$, is $\texttt{ST}(G)-\texttt{ST}(G/S)$.*

In Figure 1, the $\texttt{IM}(\{v_0\})$ is 1, while $\texttt{IM}(\{v4\}) = 16+5n+m-(7+n+m) = 9+4n$. Next, we formalize our objective.

Problem 1 (ICN: Critical Nodes as Impactful Nodes)**.** *Given $G=(V,E)$ and k, the ICN(k) problem involves identifying a set $S \subseteq V$ of size k such that $\texttt{IM}_G(S)$ is maximized.*

The reformulation of the ICN stems from the following. For any seed, its influence does not increase if some nodes from the network is removed. Larger impact indicates that each node can influence (and can be influenced by) lesser number of nodes. As a result, if S_1 and S_2 are two different sets of nodes such that $\texttt{IM}_G(S_1) < \texttt{IM}_G(S_2)$, then the influence of any seed is likely to be less (or equal) when S_2 is removed from G when compared to the case when S_1 is removed.

B. Properties of Impact

From Definition 2, one can infer that the $\texttt{IM}_G(S)$ depends on the expected influence of each vertex v in G, where the diffusion from v occurs via at least one element in S. We will prove that then $\mathcal{IM}_G$ is monotone but is neither submodular nor supermodular.

Theorem 1. *$\mathcal{IM}_G$ is monotonically increasing.*

Proof. Let S be a set of nodes. Recall that

$$\texttt{IM}_G(S) = \texttt{ST}(G) - \texttt{ST}(G/S) = \sum_{v\in V}\sigma_G(v) - \sum_{v\in V}\sigma_{G/S}(v)$$

When the probabilities are 1, $\sigma_G(v)$ is precisely the number of nodes reachable from v in G. If a node u is reachable from v only via a node from S, then u is not reachable from v in the graph G/S. Thus, $\forall v \in V : \sigma_G(v) - \sigma_{G/S}(v)$ is the number of nodes reachable from v *only* through some nodes in S. Therefore, $\forall S_1, S_2 \subseteq V : S_1 \subseteq S_2 \;\Rightarrow$

$$\begin{aligned} &\forall v \in V : (\sigma_G(v) - \sigma_{G/S_1}(v)) \leq (\sigma_G(v) - \sigma_{G/S_2}(v)) \\ \Rightarrow \quad &\sum_{v\in V}(\sigma_G(v) - \sigma_{G/S_1}(v)) \leq \sum_{v\in V}(\sigma_G(v) - \sigma_{G/S_2}(v)) \end{aligned}$$

□

The above proof assumed that the edge probabilities are all 1. The general case follows the strategy presented in [14]. Consider the sample space in which each sample point is a sub graph of G that is formed as follows: For each edge e, keep in the graph with probability p_e. Suppose that $G_1, G_2, \cdots G_\ell$ are all the sample points in the sample space. Now $\sigma(S)$ is precisely $\sum Reach(S, G_i) \times \Pr[G_i]$, where $Reach(S, G_i)$ denotes the number of nodes reachable from S in the graph G_i, and $\Pr[G_i]$ is the probability that the graph G_i is obtained by the above probabilistic process; monotonicity of reachability leads to the monotonicity of $\sigma(\cdot)$. This validates the above proof in general case.

We next establish $\mathcal{IM}_G$ is neither submodular nor supermodular. Submodularity (supermodularity) of a function is defined in terms of the marginal gain for the function. In our context, let S be a set and $v \notin S$ be a node, then the marginal gain in terms of $\mathcal{IM}_G$ is defined as follows:

$$\texttt{imgain}_G(S, v) = \texttt{IM}_G(S \cup \{v\}) - \texttt{IM}_G(S)$$

Submodularity of $\mathcal{IM}_G$ requires for all S_1, S_2 and $v \notin S_2$, $S_1 \subseteq S_2$ implies $\texttt{imgain}_G(S_1, v) \geq \texttt{imgain}_G(S_2, v)$. Conversely, for supermodularity, it is required to satisfy $\texttt{imgain}_G(S_1, v) \leq \texttt{imgain}_G(S_2, v)$.

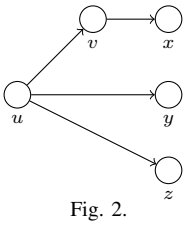


Fig. 2.

Theorem 2. *$\mathcal{IM}_G$ is not supermodular.*

Proof. Consider the G (Figure 2) where the probability associated with each edge is 1. For proving our claim, we need to show that there exists S_1, S_2 and v such that $S_1 \subseteq S_2$, $v \notin S_2$ and $\texttt{imgain}_G(S_1, v) > \texttt{imgain}_G(S_2, v)$.

Note that $\texttt{ST}(G) = 10$. Let S_1 be $\{y\}$, S_2 be $\{x, y\}$. Then, $\texttt{ST}(G/S_1) = 8$, $\texttt{IM}_G(S_1) = \texttt{ST}(G) - \texttt{ST}(G/S_1) = 2$, $\texttt{ST}(G/S_1 \cup \{v\}) = 4$, $\texttt{IM}_G(S_1 \cup \{v\}) = 6$ and $\texttt{imgain}_G(S_1, v) = 6 - 2 = 4$.

Proceeding further, $\texttt{IM}_G(S_2) = 5$ and $\texttt{IM}_G(S_2 \cup \{v\}) = 7$, and therefore, $\texttt{imgain}_G(S_2, v) = 2 < \texttt{imgain}_G(S_1, v)$. □

Theorem 3. *$\mathcal{IM}_G$ is not submodular.*

Proof. Consider the G (Figure 3) where the probability associated with each edge 1. For proving our claim, we need to show that there exists S_1, S_2 and v such that $S_1 \subseteq S_2$, $v \notin S_2$ and $\texttt{imgain}_G(S_1, v) < \texttt{imgain}_G(S_2, v)$.

Note that, $\texttt{ST}(G) = 11$. Let S_1 be $\{y\}$ and S_2 be $\{y, z\}$. Then, $\texttt{imgain}_G(S_1, v) = \texttt{IM}_G(S_1 \cup \{v\}) - \texttt{IM}_G(S_1) = 3$ and $\texttt{imgain}_G(S_2, v) = \texttt{IM}_G(S_2 \cup \{v\}) - \texttt{IM}_G(S_2) = 4$. Therefore, $\texttt{imgain}_G(S_1, v) < \texttt{imgain}_G(S_2, v)$. □

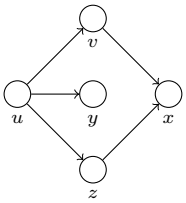


Fig. 3.

Interestingly, if the network G has the property that there is at most one path between any two nodes, then $\mathcal{IM}_G$ is submodular.

Theorem 4. *$\mathcal{IM}_G$ is submodular if there is at most one path between, any two nodes in G.*

Proof. Again, we assume that all the edge probabilities are 1. The general case follows as per the arguments presented after proof of Theorem 1. We need to prove that for any S_1, S_2 and v,

$$S_1 \subseteq S_2 \ \wedge \ v \notin S_2 \Rightarrow \texttt{imgain}_G(S_1, v) \geq \texttt{imgain}_G(S_2, v)$$

Recall $\texttt{imgain}_G(S_1, v)$ is

$$\texttt{IM}_G(S_1 \cup \{v\}) - \texttt{IM}_G(S_1) = \texttt{ST}(G/S_1) - \texttt{ST}(G/(S_1 \cup \{v\}))$$

That is, $\texttt{imgain}_G(S_1, v)$ is the number of nodes that are reachable from v and are not reachable from S_1. If any of the elements can reach v, then $\texttt{imgain}_G(S_1, v) = 0$, as there is at most one path between any two nodes in the network.

Next, for any S_2 such that $S_1 \subseteq S_2$, there are three possibilities in which elements in $S_2 - S_1$ can be selected. (a) there are some elements in $S_2 - S_1$, that can reach v, in which case, $\texttt{imgain}_G(S_2, v) = 0$; (b) None of the elements in $S_2 - S_1$ are reachable from v, in which case, $\texttt{imgain}_G(S_2, v) = \texttt{imgain}_G(S_1, v)$; (c) Some of the elements in $S_2 - S_1$ that are reachable from v, in which case $\texttt{imgain}_G(S_2, v) < \texttt{imgain}_G(S_1, v)$. □

Theorem 5. *ICN(k) problem (See Problem 1) is NP-Hard.*

Proof. In [25], Yannakakis proved that the the problem of removing an optimal number of nodes from a graph resulting in subgraph satisfying some hereditary property is NP-Hard. Hereditary property of a graph is one that is preserved in all induced subgraphs.

Consider the decision version of $ICN(k)$: does there exists set S of k nodes, whose removal from the graph G results in $\texttt{ST}(G/S) \leq T$? The strength of a graph is an hereditary property, i.e., $\texttt{ST}(G) \leq A \ \Rightarrow \texttt{ST}(G/S') \leq A$ for all $S' \subseteq V$, where V is the set of nodes in G. Therefore, our decision problem is a member of the NP-Hard class of node removal problems. □

IV. Algorithm for Finding Critical Nodes

Given the hardness of the ICN(k) problem, we will present an effective and efficient greedy heuristic strategy for identifying the critical nodes, which (likely) maximizes impact. Figure 4 presents the basic steps necessary to solve $ICN(k)$ based on this strategy.

Input: Network $G = (V, E)$ and k
Output: $S \subseteq V$
$S = \emptyset$
while $|S| < k$ **do**
  $w = \texttt{argmax}_{v \in V}\ \texttt{imgain}_G(v, S)$
  $S = S \cup \{w\}$
end while
return S

Fig. 4. Greedy Computation

The algorithm incrementally computes the set (of size k) of nodes with maximal impact; at each iteration, identifying the node that results in maximal marginal gain in impact with respect to the set computed in the previous iteration.

Note that, the maximal marginal gain computation at each step for each node (yet to be considered in S) is an expensive process. In [3], the authors presented random reachable set based efficient implementation for computing marginal gains in the context of influence maximization problem. We will employ the same implementation strategy for impact computation. We first present a short description of the strategy as presented by Borgs et al. [3].

Given a network G, let G^r is the same network with the edges reversed. A set $RR = \{G_1^r, G_2^r, \ldots, G_N^r\}$ of graphs is constructed as follows. For each G_i^r, randomly pick a node v in G^r and conduct a random walk in G^r (using the edge probabilities) starting from v. Borgs et al. proved that if a vertex v belongs to M number of elements in RR, then expected influence of v can be estimated as $\hat{\sigma}(v) = (M/N) \times |V|$. It follows from Chernoff bounds that $\hat{\sigma}$ approximates σ with relative error ϵ when $N = O(|V|/\epsilon^2)$. The marginal gain in influence due to a vertex v with respect to some set S, therefore, can computed by considering the number of RR elements which contains v but none of the elements of S. Incrementally computing marginal gain can be easily realized as follows: at each iteration identify the vertex with maximal coverage of (existing) RR set and remove all the RR elements that vertex covers before proceeding to the next iteration.

In the following, we will present the strategy that we use to compute the marginal gain in impact due to a vertex with respect to a given set using random reachable set.

A. Impact Computation using Random Reachability

In our context, we need to compute the impact of a set S, which involves computing $\sigma_G(v) - \sigma_{G/S}(v)$ for all nodes v. Let $M_{\overline{S}}$ indicate the number of elements in RR set that contains v such that there is at least one path to v independent of any node in S. Conversely, M_S indicate the number of elements in RR set that contains v such that all paths to v involve some node in S. Therefore, $\sigma_{G/S}(v)$

$= \sum_{u \in V} Pr(\exists u : v \text{ influences } u \text{ without involving any } w \in S)$
$= \sum_{u \in V} Pr(\exists u : u \text{ reaches } v \text{ in } G^r \text{ without involving any } w \in S)$
$= |V| \times$
$Pr(\exists u : u \text{ reaches } v \text{ in } G^r \text{ without involving any } w \in S)$

That is, $\hat{\sigma}_{G/S}(v) = |V| \times M_{\overline{S}}/N$. Proceeding further,

$$\hat{\sigma}_G(v) - \hat{\sigma}_{G/S}(v) = |V|/N \times (M - M_{\overline{S}}) = |V| \times M_S/N$$

Recall that, $\texttt{IM}_G(S) = \sum_{v \in V} \sigma_G(v) - \sum_{v \in V} \sigma_{G/S}(v)$. Therefore, $\texttt{IM}_G(S)$ can be estimated by counting the number of times each node in G is reachable in graphs in RR set where the reachability requires the existence of some node in S.

B. Incremental Computation of Marginal Gain in Impact

Recall that the marginal gain in impact due to a node v with respect to S is $\texttt{imgain}_G(v, S) = \texttt{IM}_G(S \cup \{v\}) - \texttt{IM}_G(S)$. Computing $\texttt{IM}_G(S)$ involves computing $|V| \times M_S^u/N$ for all $u \in V$ (let M_S^u denote the number of graphs in RR set where reachability of u requires some element in S).

That is, $\texttt{imgain}_G(v, S) = |V|/N \sum_{u \in V} \left[M^u_{S \cup \{v\}} - M^u_S \right]$. Proceeding further, $M^u_{S \cup \{v\}} - M^u_S$ is equal to the difference between number of graphs in RR where reachability of u involves v or some elements in S and number of graphs in RR where reachability of u involves some elements in S. Therefore, $M^u_{S \cup \{v\}} - M^u_S$ is the number of graphs in RR where reachability of u involves v and does not involve any element from S. Incremental computation of $imgain_G(v, S)$ (and avoid computing $\texttt{IM}_G(S \cup \{v\})$) is realized as follows. Once $\texttt{IM}_G(S)$ is computed using RR set, we remove all elements of S from each $G_i^r \in RR$. After removal, $|V| \times M_v^u/N$ for all $u \in V$ is equal to $M^u_{S \cup \{v\}} - M^u_S$, which, in turn, results in incremental computation of $\texttt{imgain}_G(v, S)$. Figure 5 outlines the method using reachable sets.

Input: $RR = \{G_1^r, G_2^r, \ldots, G_N^r\}$ and k
Output: $S \subseteq V$
$S = \emptyset$
while $|S| < k$ **do**
  $w = \texttt{argmax}_{v \in V} \sum_{u \in V} M_v^u$
  $S = S \cup \{w\}$
  Remove w from RR graphs
end while
return S

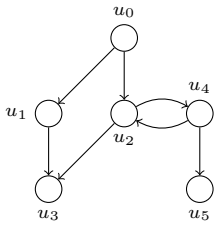
Fig. 5. Greedy with Random Reachability

C. Efficient Implementation of Incremental Computation

For the incremental computation one needs to perform reachability on each graphs in RR set in every iteration. We develop a data structure that succinctly captures the reachability information in each graphs of RR set and present effective algorithms to minimize the re-computation of reachability.

For each node $v \in V$ and for each graph G_i^r in RR set, we maintain a set $\texttt{dependOn}(v, i) \subseteq V$. The set contains the nodes such that their reachability requires v in G_i^r. If U is the set of nodes in G_i^r, then $\texttt{dependOn}(v, i)$ can be computed by subtracting from U the nodes that are reachable in G_i^r after removing v. The impact of v proportional to $\sum_{i=1}^{N} \texttt{dependOn}(v, i)$ (equal to $\sum_{u \in V} M_v^u$).

In order to facilitate incremental computation of marginal gain of impact, $imgain$, the $\texttt{dependOn}(w, i)$ must be updated for all $w \in V$ and $i \in [1, N]$ once a node $v \neq w$ with the highest impact is selected to be part of the solution. Incrementality requires the removal of v and recomputation of reachability in G_i^r. This repeated reachability can be avoided by the following update operation on $\texttt{dependOn}(w,)$.



If $u \in \texttt{dependOn}(v, i)$ then remove u from all $\texttt{dependOn}(w, i)$ $(w \neq v)$. This is because v in G_i^r impacts u (removing v will make u unreachable in G_i^r); reachability of u cannot be any more falsified (impacted) by further considering w. This is illustrated in the example figure G_i^r. The corresponding $\texttt{dependOn}$ is represented using as matrix, where the first column represent the input and each cell (r, c) is set to 1, if the c-th element is present in the $\texttt{dependOn}$ of r-th element.

	u_0	u_1	u_2	u_3	u_4	u_5
u_0	1	1	1	1	1	1
u_1		1				
$\boldsymbol{u_2}$			**1**		**1**	**1**
u_3				1		
u_4					1	1
u_5						1

If u_2 is selected as the one with the highest impact[1], then row corresponding to u_2, representing the set $\texttt{dependOn}(u_2, i)$, will be rendered unreachable in G_i^r by the removal of u_2.

Secondly, subsequent computation of impact of nodes u_0, u_1 and u_3 should not consider the unreachable nodes (u_2, u_4 and u_5), and hence, their entries (if present) are removed from the $\texttt{dependsOn}$ of u_0, u_1 and u_3.

Finally, after removal of u_2, the impact of some nodes may improve as well. Such nodes are the ones whose reachability does not depend on u_2 and which, if removed in the absence of u_2, may render some other nodes un-reachable. For instance, in the absence of u_2, removal of u_1 will render u_3 unreachable.

	u_0	u_1	u_3
u_0	1	1	1
u_1		1	1
u_3			1

Therefore, $\texttt{dependOn}$ of u_1 includes u_3 after removal of u_2. Such update is realized by *only* re-computing the $\texttt{dependOn}$ relationship of all nodes that do not belong to the $\texttt{dependOn}$ relation of the node being removed (u_2 in our example).

V. Experimental Evaluation

[1]Note that the above simply illustrates one of the N random graphs in RR. Impact of a node based on the sum of its impact in all the N elements.

The primary objective of our experiments is to evaluate the quality of the results obtained by removing certain number (referred to as *budget*) of critical nodes. We refer to our proposed method as CRIT-SET. To measure advantages of using our method, we developed two other methods, which are obvious and immediate choices for disrupting diffusion: (a) one based on removing the top k most influential nodes (TOP-INFL) and (b) one based on removing the top k most critical nodes (TOP-CRIT). We will use TOP-INFL as the baseline method.

Network-name	# Nodes	# Edges
condensed-Matter-Collab	23,133	93,497
soc-sign-Epinion	131,828	841,372
com-DBLP	317,080	1,049,866
DBLP-Tang	613,586	1,990,159
Web-Google	875,713	4,332,051

TABLE I
DATASET

We use several networks from http://snap.stanford.edu/data/. In all the experiments, following the prior works[2], we chose $p_{uv} = 1/d_{in}(v)$, where $d_{in}(v)$ is the indegree of v. The size of RR is computed based on the chosen $\epsilon = 0.5$.

We observe that the quality of the results does not improve much for smaller values of ϵ. Table I presents the basic information about the networks used in the experiments[3].

A. Criticality-indicator & Importance of Critical Nodes

This set of experiments is directed to validate the claim that that removing critical nodes indeed reduces the possible diffusion from a seed. The experiments are set up in two different ways: one focuses on removing nodes in a given influence graph induced by a seed; and the other focuses on removing nodes that reduces the influence of a given seed.

Setup I: Influence-Graph Context. For each network, we identified (using [3]) the best influential seed set of different sizes. We then use a random diffusion from that seed set to generate the influence graph–the graph where all nodes are influenced. Assuming this influence graph to be the input (that is, the objective is to maximally disrupt diffusion in this influence graph), we conduct experiments to find the impact of removing k nodes in the influence graph.

Table II presents a subset of results obtained in this experiment. The first column is the size of the influence graph generated by seed of size k (second column). The budget indicates the number of nodes to be removed. The New-infl columns indicate the influence in the input graph (after nodes are removed). We use the same size for seed set and construct them by considering the objective of maximizing its influence on $x\%(x \in [20, 90])$ of the network. We report (New-Infl) the average influence size in the network using these different seeds after the nodes are removed. It also includes the (average) percentage improvement over the baseline TOP-INFL method. The timing results are given in seconds.

I-Graph	Seed	Budget	TOP-INFL		TOP-CRIT		CRIT-SET	
			New Infl	Time (sec)	New Infl	Time (sec)	New Infl	Time (sec)
soc-sign-Epinion								
9315	10	5	3134	0.01	3015 (4%)	0.24	2857 (9%)	0.98
		10	1209	0.01	1033 (11%)	0.24	871 (25%)	1.03
		15	1121	0.01	925 (15%)	0.24	710 (35%)	1.06
		20	987	0.01	811 (17%)	0.24	629 (34%)	1.09
Network com-DBLP								
18298	10	5	4669	0.03	4587 (1%)	0.70	4017 (12%)	3.18
		10	3209	0.03	2608 (17%)	0.73	2807 **(12%)**	3.32
		15	2349	0.04	2128 (9%)	0.69	1898 (19%)	3.34
		20	2221	0.03	2066 (7%)	0.70	1898 (14%)	3.33

TABLE II
CRITICALITY-INDICATOR & IMPORTANCE WITH SETUP I

Infl Size	Seed	Budget	TOP-INFL		TOP-CRIT		CRIT-SET	
			New Infl	Time (sec)	New Infl	Time (sec)	New Infl	Time (sec)
Network DBLP-Tang								
23702	10	5	18201	186	17560 (4%)	968	16906 (7%)	1000
		10	17893	175	16755 (6%)	919	15009 (16%)	972
		15	17889	186	15313 (14%)	988	13920 (22%)	1059
		20	17264	185	14529 (16%)	961	12958 (24%)	1060
Network Web-Google								
4750	10	5	4016	421	3791 (6%)	1347	3794 (6%)	1357
		10	3840	466	3536 (8%)	1416	3524 (8%)	1441
		15	3671	455	3278 (11%)	1428	3213 (12%)	1475
		20	3494	450	2982 (15%)	1415	2919 (16%)	1466

TABLE III
CRITICALITY-INDICATOR & IMPORTANCE WITH SETUP II

Observe that, if there is a budget constraint on number of nodes that can be removed, then identifying the critical nodes can indeed save majority of the network from un-wanted diffusion. For instance, for com-DBLP network a 10-node seed can influence 18K nodes; however, removing 20 critical nodes help to reduce the result of diffusion (by virtually any 20 nodes) to around 1.9K nodes. Next observe that, in all experiments TOP-CRIT and CRIT-SET have reduced the level of diffusion more than TOP-INFL. This shows that influential nodes are not necessarily the ones that can maximize disruption in diffusion. Furthermore, reduction achieved by CRIT-SET is considerable (compared to TOP-INFL, in some case as high as 30%).

Setup II: Seed Set Context. In this setting, we considered the best seed of size k (inducing maximal influence on the given network) and focus on identifying the nodes to remove in the context of the given seed.

Recall that, we have used RR set based method. The most

[2]Probability of diffusion based on indegree is a one of the many ways to quantify the strength of nodes in spreading information—typically, referred to as the *weighted independent cascade* Our objective is not focused on the debate [1], [20] of how probability of diffusion is measured or quantified and how the efficiency of influence maximization depends on the quantification; rather our focus is to validate our characterization of criticality in terms of impact and not the efficiency of general diffusion problem. In fact, any of the efficient and effective diffusion algorithms can be used in our implementation framework. We chose the basic random reachable set based method, which is at the core of notable efficient algorithms (e.g., [23]).

[3]For lack of space, in the following sections, we present results on specific networks; however, similar results are obtained for the networks not illustrated in tables and figures.

influential nodes are the ones that has the maximal coverage of RR set ([3]). For incorporating context information, in TOP-INFL, we identify the elements (set $\mathcal{E} \subseteq$ RR) of RR set covered by the given seed, and then identify the nodes (other than those in the seed) that maximally covers $\mathcal{E}$. Similarly, for TOP-CRIT and CRIT-SET, we only consider the set $\mathcal{E}$ to compute the impact/criticality of the nodes. A node in a particular $G_i \in \mathcal{E}$ (recall, there are N number of G_i's generated by reverse random reachability) is critical only when the removal of the node disrupts the reachability to all elements of the seed in G_i. The criticality of a node is equal to the number of elements in $\mathcal{E}$, where the node is deemed critical. In other words, in all three methods, the seed nodes drive the selection of nodes to be removed. After removal of nodes, we re-compute the expected influence of the original seed set.

In Table III, we report results of experiments using setup II on some additional networks. It is important to note that in setup I, the input network is the influence graph generated by one simulation of the given network from the max seed, and all results (i.e., New Infl for each method) are with respect to the influence graph; on the other hand, in setup II, the input network is the given network, and all results (New infl for each method) are with respect to the given network. The results obtained from setup I and setup II are not comparable. However, the application of criticality in both setups presents one important commonality. The methods TOP-CRIT and CRIT-SET render better results when compared to TOP-INFL; CRIT-SET is considerably more expensive due to submodularity, which requires some re-computation of criticality of nodes.

B. Role of Budget on Node Removal

Our next set of experiments analyze the relationship between budget (number of nodes to remove) and the node-removal strategy. In particular, we are interested in understanding the difference in quality of results obtained by CRIT-SET and TOP-INFL as the budget increases.

We consider the condensed-matter-collab network (see Table I). We find k nodes to remove using CRIT-SET and TOP-INFL. We record the number of common nodes being removed (intersection size).

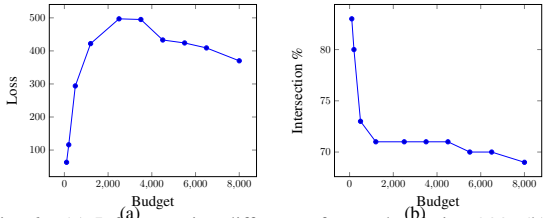


Fig. 6. (a) Influence size difference for seeds of size 300, (b) Intersection size of nodes being removed by CRIT-SET and TOP-INFL for budget range $[100 - 8,000]$.

After removal of k nodes, we consider M size seed set to start and compute the level of diffusion. Different types of seeds are computed by considering it maximal influence on $x\%$ of the network ($x \in [20\%, 90\%]$). For each seed, the influence size is computed. The average difference between the influence sizes (after removal of nodes using CRIT-SET and TOP-INFL) is recorded.

Experiment is conducted by varying k starting from 100 for two different values of M equals to 300 and 1500. Figure 6(a, b) presents the difference and intersection size against the budget values for $M = 300$. Note that as the budget increases, the difference in the influence size increases rapidly and then plateaus, and finally decreases. On the other hand, as the budget increases, the intersection size of the nodes to be removed by two methods decreases and then flattens. The observations can be explained as follows. For smaller budget k, in this case study, the intersection is high because highly critical nodes are also highly influential nodes. As a result, the difference in the influence size after removal of nodes using the two methods is not large. However, with the increase in the budget k, the methods proceed to identify moderately critical nodes (CRIT-SET) and moderately influential nodes (TOP-INFL)–these sets are not likely to be same/similar. In other words, CRIT-SET decides to remove nodes (critical nodes) that are markedly different from the nodes (influential nodes) being removed by TOP-INFL. This, coupled with the fact that removal of critical nodes disrupts the diffusion more than the removal of influential ones (as observed in the last subsection), the difference between the influence sizes after removal of influential nodes and after removal of critical nodes increases as the budget increases. The pattern continues up to certain budget after which the nodes to be removed again exhibit the same level of criticality and influence, at which point, the difference between influence size flattens and starts decreasing. This is because all the critical and influential nodes, which have some significant effect on diffusion, have been already considered for removal.

C. Critical Nodes Removal and Maximal Influence

Our next experiment focuses on validating that maximum influence achievable is significantly reduced after removal of nodes following CRIT-SET. We consider seed set of size 300 in the condensed-Matter-Collab-Network. We identify the seeds that can induce the maximal diffusion after removal of nodes, and report the number by which diffusion after the critical node removal is less than that after the influential node removal (Figure 7). CRIT-SET always outperforms TOP-INFL as the budget increases; the difference increases till the budget for removal is $2,000$. This is exactly the same pattern we observed in the last experimental setup; however, there is an important distinction between the two experiments. In the last experiment, the same seed set is used after the node-removal using CRIT-SET and TOP-INFL; in the current experiment, the best seeds (inducing maximal diffusion) is considered after removal of nodes. As a result, the seeds being considered after removal

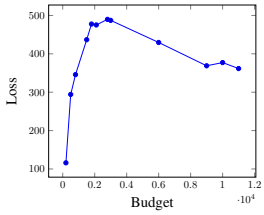


Fig. 7. Disruption of Diffusion from the Best Seed

of nodes using CRIT-SET is different from the one being considered after removal of nodes using TOP-INFL. The observation validates the claim that the maximal diffusion achievable after critical node removal is less than that achievable after influential node removal; in other words, removing critical nodes disrupts the diffusion possible from the best seeds.

D. Importance of Submodularity: TOP-CRIT *vs.* CRIT-SET

Recall that in our method, CRIT-SET, as we add a node to the solution set, we re-compute criticality of the rest of the nodes (hence, the submodularity). The TOP-CRIT, on the other hand, does not perform such re-computation and assumes that criticality of a node is a modular property. Our final set of experiments focus on the importance of submodularity in the computation of critical set of nodes (in particular, when the source of influence may not be known a priori). We compute the solution (say, A) using CRIT-SET and the solution (say, B) using TOP-CRIT. Then we compute the set of nodes R^{A-B} that are likely (probabilistic reachability) to reach the set of nodes $A - B$ (nodes that are present in A and absent in B). Similarly, we compute the set R^{B-A}. Intuitively, R^{A-B} (resp. R^{B-A}) indicates the set of nodes whose influence in the network is likely to be disrupted due to the removal of nodes in $A - B$ (resp. $B - A$). Proceeding further, if the expected influence of R^{A-B} is larger compared to that of R^{B-A}, then we claim that removing nodes in A (CRIT-SET) is likely to be more disruptive than removing nodes in B (TOP-CRIT). Table IV presents the results of our experiments and affirms the importance of submodularity in CRIT-SET method.

Budget	$\sigma(R^{A-B})$	$\sigma(R^{B-A})$
Network DBLP-Tang		
5	74191	3974
10	33453	1635
15	32668	10245
20	91020	15430

TABLE IV
IMPORTANCE OF SUBMODULARITY

VI. CONCLUSION

We study the problem of identifying critical nodes for disrupting influence in complex network under IC diffusion model. We introduce the characterization of criticality in terms of impact, which, in turn, describes the reduction in the diffusion strength of the network. We present a greedy heuristics for impact computation and design experiments to validate the effectiveness of our characterization in addressing the problem.

We plan to consider different heuristics and implementation strategies to realize the computation of impact; the goal being application to very large networks efficiently without compromising the quality. Another avenue of research along this line of work, includes associating costs and hard constraints on the nodes (e.g., some nodes may not be removed, some nodes may incur prohibitive cost to remove) and addressing the problem of constrained cost-effective disruption.

REFERENCES

[1] A. Arora, S. Galhotra, and S. Ranu. Debunking the myths of influence maximization: An in-depth benchmarking study. In *ACM International Conference on Management of Data*, pages 651–666, New York, NY, USA, 2017. ACM.

[2] J. Aspnes, K. L. Chang, and A. Yampolskiy. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, Vancouver, British Columbia, Canada, January 23-25, 2005*, pages 43–52, 2005.

[3] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In *Proc. of the 25th SODA 2014*, pages 946–957, 2014.

[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *ACM SIGKDD*, pages 1029–1038, 2010.

[5] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *ACM SIGKDD*, pages 199–208, 2009.

[6] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *Proc. of the 10th ICDM 2010*, pages 88–97, 2010.

[7] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng. Staticgreedy: Solving the scalability-accuracy dilemma in influence maximization. In *ACM International Conference on Information & Knowledge Management*, pages 509–518, 2013.

[8] S. Galhotra, A. Arora, and S. Roy. Holistic influence maximization: Combining scalability and efficiency with opinion-aware models. In *International Conference on Management of Data*, pages 743–758. ACM, 2016.

[9] A. Goyal, F. Bonchi, and L. Lakshmanan. A data-based approach to social influence maximization. *Proc. VLDB Endow.*, 5(1):73–84, 2011.

[10] A. Goyal, W. Lu, and L. Lakshmanan. Celf++: Optimizing the greedy algorithm for influence maximization in social networks. In *Proc. of the 20th WWW*, pages 47–48. ACM, 2011.

[11] X. He, G. Song, W. Chen, and Q. Jiang. Influence blocking maximization in social networks under the competitive linear threshold model technical report. *CoRR*, 2011.

[12] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou. Rumor source identification in social networks with time-varying topology. *IEEE Transactions on Dependable and Secure Computing*, 15(1):166–179, 2018.

[13] K. Jung, W. Heo, and W. Chen. IRIE: scalable and robust influence maximization in social networks. In *12th ICDM 2012.*, pages 918–923, 2012.

[14] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *ACM SIGKDD*, pages 137–146, 2003.

[15] E. B. Khalil, B. Dilkina, and L. Song. Scalable diffusion-aware optimization of network topology. In *ACM SIGKDD*, pages 1226–1235, 2014.

[16] J. Kleinberg. *Cascading Behavior in Networks: Algorithmic and Economic Issues*, chapter 24. Cambridge University Press, 2007.

[17] C. J. Kuhlman, V. S. Anil Kumar, M. V. Marathe, S. S. Ravi, and D. J. Rosenkrantz. Finding critical nodes for inhibiting diffusion of complex contagions in social networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 111–127, 2010.

[18] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *ACM SIGKDD*, pages 1059–1068, 2010.

[19] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *ACM SIGKDD*, pages 420–429, 2007.

[20] W. Lu, X. Xiao, A. Goyal, K. Huang, and L. V. S. Lakshmanan. Refutations on "debunking the myths of influence maximization: An in-depth benchmarking study". *CoRR*, 2017.

[21] N. Ohsaka, T. Akiba, Y. Yoshida, and K. I. Kawarabayashi. Fast and accurate influence maximization on large networks with pruned monte-carlo simulations. In *Proceedings of the AAAI*, pages 138–144, 2014.

[22] W. Pullan. Heuristic identification of critical nodes in sparse real-world graphs. *Journal of Heuristics*, 21:577–598, 2015.

[23] Y. Tang, Y. Shi, and X. Xiao. Influence maximization in near-linear time: A martingale approach. In *SIGMOD*, pages 1539–1554, 2015.

[24] M. Ventresca and D. Aleman. Efficiently identifying critical nodes in large complex networks. *Computational Social Networks*, 2, 2015.

[25] M. Yannakakis. Node-and edge-deletion np-complete problems. In *ACM Symposium on Theory of Computing*, pages 253–264, 1978.

[26] W. Zang, P. Zhang, C. Zhou, and L. Guo. Discovering multiple diffusion source nodes in social networks. *Procedia Computer Science*, 29:443 – 452, 2014.