IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 22, 2016, accepted January 17, 2017, date of publication March 1, 2017, date of current version August 29, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2672666

Transitivity Demolition and the Fall of Social Networks

HUNG T. NGUYEN[1], NAM P. NGUYEN[2], TAM VU[3], (Member, IEEE), HUAN X. HOANG[4], AND THANG N. DINH[1], (Member, IEEE)

[1]Computer Science Department, Virginia Commonwealth University, Richmond, VA 23220 USA
[2]Computer and Information Sciences Department, Towson University, Towson, MD 21252 USA
[3]Computer Science and Engineering Department, University of Colorado, Denver, CO 80204 USA
[4]Information Technology Department, Vietnam National University, Hanoi 10000, Vietnam

Corresponding author: Nam P. Nguyen (npnguyen@towson.edu)

This work was supported in part by the Google Faculty Awards 2013-R2-634, and in part by the National Science Foundation under Grants CNS 1452628 and SCH-1602428. The work of N. P. Nguyen was supported in part by the NVIDIA Corporation Hardware Grant, and in part by The Jess and Mildred Fisher Endowed Professor award, Towson University.

ABSTRACT In this paper, we study crucial elements of a complex network, namely its nodes and connections, which play a key role in maintaining the network's structure and function under unexpected structural perturbations of nodes and edges removal. Specifically, we want to identify vital nodes and edges whose failure (either random or intentional) will break *the most number of connected triples (or triangles)* in the network. This problem is extremely important, because connected triples form the foundation of strong connections in many real-world systems, such as mutual relationships in social networks, reliable data transmission in communication networks, and stable routing strategies in mobile networks. Disconnected triples, analog to broken mutual connections, can greatly affect the network's structure and disrupt its normal function, which can further lead to the corruption of the entire system. The analysis of such crucial elements will shed light on key factors behind the resilience and robustness of many complex systems in practice. We formulate the analysis under multiple optimization problems and show their intractability. We next propose efficient approximation algorithms, namely, DAK-n and DAK-e, which guarantee an $(1 - 1/e)$-approximate ratio (compared with the overall optimal solutions) while having the same time complexity as the best triangle counting and listing algorithm on power-law networks. This advantage makes our algorithms scale extremely well even for very large networks. In an application perspective, we perform comprehensive experiments on real social traces with millions of nodes and billions of edges. Empirical results indicate that our approaches achieve comparably better solution quality while are up to $100\times$ faster than the current state-of-the-art methods.

INDEX TERMS Triangle breaking, social networks, approximation algorithms.

I. INTRODUCTION

Resilience to unexpected perturbations is perhaps one of the most desirable properties for real-world complex systems, such as the World Wide Web, communication networks, transportation networks, biological networks and social information networks. In general, the resilience of a network evaluates how much the network's normal function is affected in case of external or undesired perturbation, i.e., it measures the network in response to unexpected events such as adversarial attacks and random failures [1]. In order to improve the robustness of real-world systems, it is therefore important to obtain key insights into the structural vulnerabilities of the networks representing them. A major aspect of this is to analyze and understand the effect of failure (either intentionally or at random) of individual components on the degree of clustering in the network.

Clustering, or more particularly, *the number of connected triples/triangles*, is a fundamental network property that has been shown to be relevant to a variety of topics, such as clusters of genes in biological networks, forwarding and routing tables mobile networks, and especially strong connection of communities users in online social networks (OSNs) [2]–[4]. Connected triples nicely capture the social intuition "a friend of your friend is also your friend" [5], and thus, is the fundamental pattern of information diffusion in multiple systems. For example, consider the propagation of information

through a social network, such as the spread of a rumor. A growing body of work has identified the importance of the number of connected triples to such propagation; the more connected triples a network has, the easier it is for information to propagate [6]–[10]. Connected triples are also behind the fall of some online social sites, such as MySpace and Friendster, as they suffered a catastrophic degrade of active users, activity traffic, and consequently, popularity in the cyberspace. For instance, Friendster claimed to have over 100 million users at its peak, but most them had quit and fled to other networks (e.g., Facebook) by the end of 2009 [11], [12], triggering a cascade of broken bonds and friends leaving Friendster. The identification of elements that crucially affect the number of connected triples in the network, as a result, is of great impact.

The importance of connected triples is not limited to just social networks; in the context of air transportation networks, work in [13] argued that those connected triples of such a network is beneficial, as passengers for a canceled flight can be rerouted to connecting flights more easily. This metric also plays an important role in the network community structure, which is the core of mobile forwarding and routing strategies in Delay Tolerant Networks (DTNs). Particularly, [14] has shown the correlation between the number of disconnected triples and the significant degrade of forwarded packets in DTNs. In addition, as a matter of homeland security, the critical elements for clustering in homeland communication networks should receive greater resources for protection; in complement, the identification of critical elements in a social network of adversaries could potentially limit the spread of information in such a network.

Many measures have been proposed for evaluating the resilience of technological and biological systems; however, there are only few work suggested for social networks. Most studies in the literature focus on how the network behaves under perturbation using measures such as pair-wise connectivity [15], natural connectivity [16], or using centrality measures, e.g., degree, betweeness [17], the geodesic length [1], eigenvector [18], etc. Nevertheless, most of them (1) focus only on the local but not the global network' structure, and (2) do not take mutual interactions and social relationships into account. These limits drive the need for another metric for social resilience. To our knowledge, none of the existing work has examined the number of connected triples from the perspective of vulnerability - as evidenced by the examples above, the damage made by the broken triples, resulted from element-wise failures, can potentially have severe effects on the functionality of the network.

Our study in this paper investigates the structural resilience of complex networks, particularly OSNs, under the scenarios of element-wise failures due to adversary attacks or random failures. Our goal is to discover and protect critical network' elements (nodes and links) whose failures will break most triples in the network. This research largely extends our preliminary work presented in [19]. In a nutshell, our contributions are

1) We study the resilience of social networks through the number of connected triples. This an important structural vulnerability of an OSN that can greatly affect its popularity among the crowds. We formulate the analysis under multiple optimization problems, and show their hardness and intractability.
2) We propose efficient approximation algorithms to identify triangle-breaking points (i.e., nodes and links) in the network structure: DAK-n algorithm for node removal and DAK-e algorithm for edge removal. Our proposed approaches guarantee are a small constant factor in comparision to optimal solutions. Interestingly, both DAK-n and DAK-e have the same time complexity with the best triangle counting/listing algorithms, $O(m^{\frac{3}{2}})$. This makes our algorithms scale extremely well for large social data.
3) We also investigate the *input-dependent bounding* technique previously appeared in [20] for influence maximization problem. The input-dependent bound usually gives better approximation guarantee than the worst-case bound since it accounts for the particular instance of the problem and particular run of the algorithms. As shown in the experiments, the input-dependent bounds vastly improve over the worst-case guarantee and in many instances our approaches produce the optimal solutions.
4) We carry out extensive experiments in comparison with state-of-the-art methods on real-world data with millions of nodes and edges. The results show that DAK-n and DAK-e substantially outperform the other methods in terms of time consumption: They are up to 100× faster than the direct competitor, GreedyAll [21], which was shown be the best available method in terms of solution quality and scalability.

Paper organization: Section II reviews studies that are related to our work. Section III describes our model and problem definitions. Sections IV shows the intractability of these investigating problems. Sections V and VI present our algorithms DAK-n and DAK-e for the problems of interested in terms of node and edge detection, respectively. In section VII, we report empirical results of our approaches in comparison with other strategies. Finally, section VIII concludes the paper.

II. RELATED WORK

Many metrics and approaches have been proposed to account for network robustness and vulnerability [22]–[26]. While each of these measures has its own emphasis and rationality, they often come with several shortcomings that prevent them from capturing desired characteristics of network connectivity and resilience. For example, measures based on shortest path are rather sensitive to small changes (e.g. removing edges or nodes); algebraic connectivity and diameter are not meaningful for disconnected graphs (all disconnected graphs have the same values); number of connected components

TABLE 1. List of Symbols.

Notation	Meaning
n	Number of vertices/nodes ($N = \|V\|$)
m	Number of edges/links ($M = \|E\|$)
d_u	The degree of u
$N(u)$	The set of u's neighbors
$Tri(u)$	The set of triangles on a node u
$T(u) = \|Tri(u)\|$	The number of triangles on u
$Tri(u, v)$	The set of triangles on an edge (u, v)
$Tri(S) = \cup_{u \in S} Tri(u)$	The set of triangles on $S \subseteq V$
$Tri(F) = \cup_{(u,v) \in F} Tri(u, v)$	The set of triangles on a subset of edges $F \subseteq E$

and component sizes, arguably, do not fully reflect level of network connectivity.

Vulnerability assessment has attracted a large amount of attention from the network science community. Work in the literature can be divided into two categories: Measuring the robustness and Manipulating the robustness of a network. In measuring the robustness, different measures and metrics have been proposed such as the graph connectivity [15], the diameter, relative size of largest components, and average size of the isolated cluster [17]. Other work suggests using the minimum node/edge cut [27] or the second smallest non-zero eigenvalue or the Laplacian matrix [28]. In terms of manipulating the robustness, different strategies has been proposed such as [17] and [29], or using graph percolation [30]. Other studies focus on excluding nodes by centrality measures, such as betweeness and the geodesic length [1], eigenvector [18], the shortest path between node pairs [22], the pairwise connectivity [15], propagation of worms and cascading failures [31], [32]. More information of general vulnerability assessment can be found in [16] and references therein.

Community structure [33]–[35] is an another common pattern found in real-world networks. Network structural vulnerability in social networks, has so far been an untrodden area. In a related work [36], the authors introduced the community structure vulnerability to analyze how the communities are affected when top k vertices are excluded from the underlying graphs. They further provided different heuristic approaches to find those critical components in modularity-based community structure. [37] suggested a method based on the generating edges of a community to find the critical components.

Counting and listing triangles in a graph is an important problem, motivated by applications in a variety of areas. The problem of counting triangles on a graph with n vertices and m edges can be performed in a straightforward manner in $O(mn)$. This has been improved to $O(m^{3/2})$ in [38] and $O(m^{\frac{2w}{w+1}})$ where $w < 2.376$ is the exponent of matrix multiplication [39]. To improve the performance of triangle counting in large graphs, parallel algorithms are also studied in [40]. There are also several works on approximate triangle counting [41]–[43]. Recently, the k-triangle-breaking-node and k-triangle-breaking-edge problems are investigated in [21]. The authors provides NP-completeness proofs and greedy algorithms for the problems. Unfortunately, the NP-completeness proofs contains fundamental flaws that cannot be easily fixed.

III. MODEL AND PROBLEM DEFINITION

In this section, we first define the main problem of interest, describe its four triangle-breaking variants, and then show their NP-hardness. Based on the submodularity property of the objective functions, the approximability is stated accordingly for each problem based on the rich literature of optimizing submodular functions [44], [45].

The list of symbols is presented in Table 1.

A. MODEL

We represent a social network by an undirected graph $\mathcal{G} = (V, E)$ with $|V| = n$ nodes and $|E| = m$ undirected edges. A set of 3 nodes is call a *triple* (or *triangle*) if every pair of those nodes is connected by an edge. A triangle *breaks* if at least one node or edge is removed or excluded from the graph. Given a graph $G = (V, E)$, we investigate different models in which the adversary attempts to break the most number of triangles in the graph by removing nodes and edges either intentionally or at random. In what following, we define four variants of the triangle-breaking problem based on node and edge removals.

B. PROBLEM DEFINITION

Definition 1 (k-Triangle-Breaking-Node): Given an undirected graph $G = (V, E)$ and budget size k, find a subset S^* of k nodes whose removal will break the maximum number of triangles in G:

$$\begin{aligned} S^* = \arg\max \ & |Tri(S)| \\ \text{s.t. } & |S| \leq k \\ & S \subseteq V, \end{aligned}$$

where $Tri(S)$ is the set of triangles with at least a node in S:

$$Tri(S) = \{(u, v, w) \mid (u, v), (v, w), (w, u) \in E \text{ and } \{u, v, w\} \cap S \neq \emptyset\}.$$

k-triangle-breaking-node can also be formulated as an Integer Linear Programming problem (ILP). For each $u \in V$,

define $x_u \in \{0, 1\}$ such that

$$x_u = \begin{cases} 1 & \text{if node } u \text{ is removed} \\ 0 & \text{otherwise,} \end{cases}$$

and for each triangle $(u, v, w) \in Tri(V)$, define an integral variable $y_{uvw} \in \{0, 1\}$ that satisfies

$$y_{uvw} = \begin{cases} 1 & \text{if } (u, v, w) \text{ is broken} \\ 0 & \text{otherwise.} \end{cases}$$

Recall that k-**triangle-breaking-node**'s goal is to remove k nodes, i.e., $\sum_{u \in V} x_u \leq k$, to break the maximum number of triangles, i.e., to maximize the objective function $\sum_{(u,v,w) \in Tri(V)} y_{uvw}$. Because a triangle (u, v, w) is only broken if at least one node in $\{u, vw\}$ is chosen to be removed the following constraint is therefore imposed,

$$x_u + x_v + x_w \geq y_{uvw}.$$

The final ILP formulation of k-**triangle-breaking-node** is

$$\begin{aligned} \max \quad & \sum_{(u,v,w) \in Tri(V)} y_{uvw} \\ \text{s.t.} \quad & \sum_{v \in V} x_v \leq k, \\ & x_u + x_v + x_w \geq y_{uvw}, \quad \forall (u, v, w) \in Tri(V), \\ & x_u, y_{uvw} \in \{0, 1\}. \end{aligned} \tag{1}$$

Note that this ILP formulation forms a special case of the **Max-k-Coverage** [46] problem: Given an universe set of elements $\mathcal{U}$ and a collections of subsets of $\mathcal{U}$, $\mathcal{S} = \{S_1, \ldots, S_n\}$ where $S_i \subseteq \mathcal{U}$, the general **Max-$k$-Coverage** problem asks for k subsets of $\mathcal{S}$, $\hat{\mathcal{S}} = \{\hat{S}_1, \ldots, \hat{S}_k\}$, to maximize the coverage **Cover**$(\hat{\mathcal{S}})$ of $\hat{\mathcal{S}}$ where

$$\textbf{Cover}(\hat{\mathcal{S}}) = \left| \bigcup_{i=1}^{k} \hat{S}_k \right|,$$

is the number of distinct elements in the union of $\hat{S}_i$, $i = 1...k$. We call the number of subsets that an element appears in the *frequency* of that element. Thus, in the Eq. 1 the universe set is $\mathcal{U} = Tri(V)$ and the collection of subsets is $\mathcal{S} = \{Tri(v) \mid v \in V\}$. This special case of **Max-k-Coverage** also satisfies the condition that *all the elements have the same frequency of three* as each triangle involves exactly three nodes.

*Definition 2 (k-**Triangle-Breaking-Edge**):* Given an undirected graph $G = (V, E)$ and budget size k, find a subset F^* of k edges whose removal will break the maximum number of triangles in G:

$$\begin{aligned} F^* = \arg\max \ & |Tri(F)| \\ \text{s.t.} \ & |F| \leq k \\ & F \subseteq E, \end{aligned}$$

where $Tri(F)$ is the set of triangles with at least an edge in F:

$$\begin{aligned} Tri(F) = \{(u, v, w) \mid & (u, v), (v, w), (w, u) \in E \\ & \text{and } \{(u, v), (v, w), (w, u)\} \cap F \neq \emptyset\}. \end{aligned}$$

The equivalent ILP of k-**triangle-breaking-edge** is,

$$\begin{aligned} \max \quad & \sum_{(u,v,w) \in Tri(V)} y_{uvw} \\ \text{s.t.} \quad & \sum_{(u,v) \in E} x_{uv} \leq k, \\ & x_{uv} + x_{vw} + x_{wv} \geq y_{uvw}, \quad \forall (u, v, w) \in Tri(V), \\ & x_{uv}, y_{uvw} \in \{0, 1\}, \end{aligned} \tag{2}$$

where

$$x_{uv} = \begin{cases} 1 & \text{if edge } (u, v) \text{ is removed,} \\ 0 & \text{otherwise.} \end{cases}$$

for all $(u, v) \in E$.

k-**triangle-breaking-edge** is also forms a special case of **Max-k-Coverage** in which the elements to be covered are the triangles in G, and the collection of subsets includes the set of triangles involving each edge $(u, v) \in E$. As each triangle consists of three edges, *the frequency of each element in this instance is also three*. Moreover, any two subsets have at most one triangle in common.

We also formulate the converse variants in which we want to break a certain number (or a percentage of the total number) of triangles by removing the least number of nodes/edges from the graph. Their definitions and ILP formulations are defined in the following paragraphs

*Definition 3 (**Min-Triangle-Breaking-Node**):* Given an undirected graph $G = (V, E)$ and a positive integer $p \leq |Tri(V)|$, find a minimum-size subset S of **nodes** whose removal will break at least p triangles in G.

The ILP for **min-triangle-breaking-node**is

$$\begin{aligned} \min \quad & \sum_{v \in V} x_v \\ \text{s.t.} \quad & \sum_{(u,v,w) \in Tri(V)} y_{uvw} \geq p, \\ & x_u + x_v + x_w \geq y_{uvw}, \\ & x_u, y_{uvw} \in \{0, 1\}. \end{aligned} \tag{3}$$

*Definition 4 (**Min-Triangle-Breaking-Edge**):* Given an undirected graph $G = (V, E)$ and a positive integer $p \leq |Tri(V)|$, find the minimum-size subset F of **edges** whose removal will break at least p triangles in G.

The ILP for **min-triangle-breaking-edge**is

$$\begin{aligned} \max \quad & \sum_{(u,v) \in E} x_{uv} \\ \text{s.t.} \quad & \sum_{(u,v,w) \in Tri(V)} y_{uvw} \geq p, \\ & x_{uv} + x_{vw} + x_{wv} \geq y_{uvw}, \\ & x_{uv}, y_{uvw} \in \{0, 1\}. \end{aligned} \tag{4}$$

Note that **min-triangle-breaking-node** and **min-triangle-breaking-edge** are special cases of the **Partial Set Cover** problem [44]. The **Partial Set Cover** problem is a variation of the *set cover* problem. Given an universe set $\mathcal{U}$,

TABLE 2. Summary of Complexity and Best Approximation Guarantees.

Problem	Complexity	Best approximation ratio
k-triangle-breaking-node	NP-complete	19/27 [45]
min-triangle-breaking-node	NP-complete	3 [44]
k-triangle-breaking-edge	NP-complete	19/27 [45]
min-triangle-breaking-edge	NP-complete	3 [44]

a collection of subsets of $\mathcal{U}$, Partial Set Cover finds a sub-collection to cover only a required number p of the elements in $\mathcal{U}$. Thus, min-triangle-breaking-node and min-triangle-breaking-edge are instances of Partial Set Cover in which each element is in exactly three subsets and the intersection of any three subsets contains at most one element.

IV. HARDNESS AND APPROXIMABILITY

This section discusses the complexity and presents the currently available approximation guarantees for our proposed problems. The summary of the complexity and approximability results for the studied problems is presented in Table 2.

A. NP-COMPLETENESS

Recent work of Li et al. [21] attempted to prove the NP-completeness of problems similar to k-triangle-breaking-node and k-triangle-breaking-edge. Unfortunately, their proofs contained some flaws which are not easily addressed. Specifically, the proof of Theorem 2.1 [21] relies on a weaker constraint of the set system: ''the intersection of any *three* subsets in $\mathcal{S}$ has at most one element''. Indeed, for k-triangle-breaking-edge, the correct (and stronger) condition should be: the intersection of any *two* subsets in $\mathcal{S}$ has at most one element. Moreover, their proof relies on the assumption that if a problem is not NP-hard then there is a polynomial-time algorithm to solve it. We do not know yet if there exist NP-intermediate problems between NP and P. Consequently, the correctness of the reduction cannot be confirmed.

We show that all four aforementioned variants are indeed NP-complete problems. We present a simple NP-completeness proof of min-triangle-breaking-node (similarly k-triangle-breaking-node) via reduction from the Vertex-Cover problem [46]. The decision versions of k-triangle-breaking-node (similarly min-triangle-breaking-node) can be polynomial-time reducible from the following decision problem, called *Node-Triangle-Free*:

"Given a undirected graph $G = (V, E)$ and a number k, can we delete k nodes from G so that it is triangle-free (or in other words, there is no more triangles in G)?".

We show an important result that *Node-Triangle-Free* is polynomial-time reducible from the decision version of Vertex Cover problem (definition below).

"(Vertex Cover) Given a graph $G = (V, E)$ and an integer $0 < k < |V|$, is there a vertex-cover of size k?".

1) REDUCTION

Let $\Phi =< G = (V, E), k >$ be an instance of the vertex cover problem. For each edge $(u, v) \in E$, we add to G a new node t_{uv} and connect t_{uv} to both u and v. Let G' be the new graph. We shall reduce ϕ to an instance $\Lambda =< G', k >$ of *Node-Triangle-Free*. Obviously, if we have a vertex-cover $S \subset V$ of size k in G then we can delete the same set of nodes S in G' to obtain a triangle-free graph. In the reverse direction, we can assume without lost of generality that t_{uv} will never be removed. The reason is that we can always remove u or v and break an equal or greater number of triangle(s). Thus a subset of size k that its removal makes G' triangle-free must induce a vertex-cover of size k in G.

This reduction consequently sets forth the NP-Completeness of k-triangle-breaking-node.

Theorem 1: The problems k-triangle-breaking-node and min-triangle-breaking-node are NP-complete.

By a very similar reduction, both k-triangle-breaking-edge and min-triangle-breaking-edge can be polynomial-time reducible to the following problem:

"Can we delete k edges from a graph $G = (V, E)$ so that it is triangle-free (or in other words, there is no more triangles in G)?".

The following Theorem is obtained from [47].

Theorem 2: k-triangle-breaking-edge and min-triangle-breaking-edge are NP-complete.

B. APPROXIMABILITY

Since min-triangle-breaking-node and min-triangle-breaking-edge problems are special cases of the Partial Set Cover problem with bounded frequencies $f = 3$ [44], the primal-dual algorithm in [44] provides a 3-approximation algorithm for both problems. Instead of operating on sets, the primal-dual algorithm works on the elements in the universe set $\mathcal{U}$. It assigns a dual covering cost for each element that signifies the selection of a set to cover that element. The basic operation of the algorithm is increasing all the dual covering costs of those that have not been covered simultaneously until the total cost of uncovered elements in a set equals 1 (the cost of choosing that set). The corresponding set is then selected to the solution and the algorithm continues until satisfying the covering requirement. To achieve the f-approximation factor, the algorithm assumes that we know a set in the optimal solution (simply by trying all the possible sets) and applies the primal-dual selection on the rest. Therefore, we obtain the following result.

Theorem 3: There exist 3-approximation algorithms for min-triangle-breaking-node and min-triangle-breaking-edge.

The k-triangle-breaking-node and k-triangle-breaking-edge problems are special cases of Max-k-Coverage and the Pipage-rounding method in [45] results in an approximation algorithm with ratio $1-(1-1/3)^3 = 19/27$.

The Pipage-rounding technique is a general method providing worst-case approximation guarantees for a large class of discrete optimization problems, including Max-k-Coverage, with assignment-type constraints. It first reformulates the problem into a non-linear program which has an integral optimum and is at least $1-(1-1/f)^f$ greater than the starting problem at any feasible solution. It then finds an integral solution of the non-linear program in two phases: 1) solving the non-integral relaxation of the problem and 2) transform the non-integral solution to an integral one by pipage rounding. The relaxation is polynomially solvable and the second phase takes the solution and rounds it in the manner that the objective value of rounded solution can only increase and get closer to integral numbers. As shown in [45], each rounding circle in Pipage-rounding brings one element in the current solution to integral value. The approximation factor follows directly from the properties of the non-linear program and the rounding procedure. Therefore, we obtain the following result.

Theorem 4: There exist 19/27-approximation algorithms for k-triangle-breaking-node and k-triangle-breaking-edge.

Remarks: Both the primal-dual method in [44] and the pipage-rounding algorithm in [45] have high time complexity and are not scalable for large networks. As a result, efficient algorithms that can be applied on large-scale data are of desire. In next sessions, we propose efficient discounting algorithms for the studied problems on very large-scale networks with just a slightly looser approximation ratio.

V. ALGORITHMS FOR k-TRIANGLE-BREAKING-NODE

In this section, we first present a naive Greedy algorithm (Alg. 1) to solve k-triangle-breaking-node problem. The greedy strategy is known to obtain a $(1-1/e)$- approximate solution; however, is time consuming and therefore is prohibitive for practical large-scale data. To address the scalability issue, we propose k-triangle-breaking-node Discounting Algorithm (DAK-n - Alg. 2) which achieves the same solution quality and approximation guarantees and is at least k time faster than the naive Greedy algorithm.

Algorithm 1 Greedy Algorithm for k-**Triangle-Breaking-Node** (Simple_Greedy)

1: $S \leftarrow \emptyset$;
2: **for** i = 1 **to** k
3: $\quad S \leftarrow S + \arg\max_{v \in V \setminus S} \Delta_S(v)$;
4: **return** S

A. NAIVE GREEDY ALGORITHM

The first algorithm (Alg. 1) selects at each step the node u that breaks the most number of triangles, i.e., $u = \arg\max_{v \in V \setminus S} \Delta_S(v)$, and then adds u to the solution S. This algorithm continues until k nodes have been selected into the returned solution S.

Since k-triangle-breaking-node is a special case of Max-k-Coverage, the native greedy algorithm provides a performance guarantee of $(1-1/e)$ for k-triangle-breaking-node. Another way of proving this is to show that the main objective function (the number of broken triangles) is monotone and submodular, which in turn admits a nearly optimal greedy approximation algorithm [21].

The complexity of Alg. 1 is $O(kmn)$ assuming the budget of k nodes. In a recent work, the time complexity for Alg. 1 is brought down to $O(km^{3/2})$ in [21] using the fast triangle computation method in [38]. For large value of $k = \theta(n)$, the time-complexity of the algorithm in [21] could be as high as $O(nm^{3/2})$ which is very expensive and not scalable for practical large size data. To this end, we present in next section our scalable Discounting Algorithms for k-triangle-breaking-node with time complexity $O(m^{3/2} + km)$ which is up to $m^{1/2}$ times faster than the algorithm in [21].

B. DISCOUNTING ALGORITHM FOR k-TRIANGLE-BREAKING-NODE

Our Discounting Algorithm for k-triangle-breaking-node (DAK-n - Alg. 2) speeds up significantly the simple greedy algorithm. For small values of k, this algorithm requires as much time as the best algorithm for counting the number of triangles. The core efficiency of DAK-n is that it employs a smart updating technique to keep track of the number of effective triangles associated with each of the remaining nodes. In particular, DAK-n employs an adaptive strategy in computing the marginal gains (the number of broken triangles) when nodes are removed one after another. At each round, the node v that breaks the most number of triangles is selected into the solution. Node v is then excluded from the structure and the procedure repeats itself on the remaining nodes and *recomputes* efficiently the new marginal gain for each node u.

We structure DAK-n into two phases. The first phase (lines 1–8) extends the algorithm in [38] to compute the number of triangles that are incident with each node in the graph. This algorithm was proved to be time-optimal in $\theta(m^{3/2})$ for triangle-listing, and has been shown to be very efficient in practice. The second phase starts at line 9 where it creates a *Max-priority-queue* to ranks nodes according to values in T. DAK-n then (lines 9–18) repeats the vertex selection for k rounds. In each round, DAK-n selects a node u_{max} with the highest value of $\Delta_S(u) = T(u)$ (from top of the priority queue) into the solution. It then removes u_{max} from the graph, and performs the necessary updates on $T(u)$ for all $u \in V \setminus S$. DAK-n subsequently updates the indexes of v and w in the queue according to their new values in T.

Algorithm 2 Discounting Algorithm for k-**Triangle-Breaking-Node** (**DAK-n**)

Phase 1:

1: Number nodes from 1 to n such that $u < v$ implies $d(u) \leq d(v)$.
2: $S \leftarrow \emptyset$;
3: **for each** $u \in V$ **do** $T(u) \leftarrow 0$;
4: **for** $u \leftarrow n$ to 1 **do**
5: **for each** $v \in N(u)$ with $v < u$ **do**
6: **for each** $w \in A(u) \cap A(v)$ **do**
7: Increase $T(u)$, $T(v)$ and $T(w)$ by one;
8: Add u to $A(v)$;

Phase 2:

9: $Q \leftarrow$Max-Priority-Queue(T)
10: **for** i = 1 **to** k
11: $u_{max} = Q.pop()$;
12: Remove u_{max} from G and add u_{max} to S;
13: **for each** $v \in N(u_{max})$ **do**
14: **for each** $w \in N(v)$ **do**
15: **if** $v, w \in N(u_{max}) \setminus S$ **then**
16: Decrease $T(v)$ and $T(w)$ by one;
17: $Q.update(v, T)$;
18: $Q.update(w, T)$;
19: **return** S

The key efficiency of **DAK-n** lies in its update procedure for $\Delta_S(u) = T(u)$. Specifically, the total update for all $O(n)$ values of $\Delta_S(u)$ after removing u_{max} can be done in *linear time* as indicates in lines 15 – 18. The linear time update is made possible due to the information on the number of triangles involving each node. This significantly reduces the complexity for computing the marginal gain $\Delta_S(u)$ and speeds up the node selection process.

1) COMPLEXITY

The first phase takes $O(m^{3/2})$ as analyzed in [38]. The second phase takes a linear time in each round and has a total time complexity $O(k(m+n))$ as creating and maintaining the Max-priority queue requires $O(n \log n)$. In each sequential round, the algorithm checks all the neighbors v of u_{max} and for each neighbor, it examines all the neighbors of v. Thus, the total complexity of checking at a round is $\sum_{v \in N(u_{max})} d_v \leq 2m$ where d_v is the degree of v. Each update (Lines 17-18) takes constant time since $T(v)$ and $T(w)$ decrease by 1 and the queue Q needs to move v, w at most one level in the queue. Hence, the overall complexity is $O(m^{3/2}+km)$. For $k < m^{1/2}$, the algorithm has an effective time-complexity $O(m^{3/2})$, which is the same as the counting triangles procedure.

2) APPROXIMATION GUARANTEES

DAK-n respects the spirit of Greedy method as it selects the node with the highest marginal gain at each step. As a result, **DAK-n** retains the approximation guarantees of the greedy method for **Max-k-Coverage**. The following theorem summarizes our suggested approach.

Theorem 5: **DAK-n** algorithm is an $(1 - 1/e)$-approximation algorithm for k-**triangle-breaking-node** with a complexity of $O(m^{3/2} + km)$.

Note that the naive Greedy (Alg. 1) and Discounting Algorithms (Alg. 2) can be easily adapted for **min-triangle-breaking-node** by terminating node selection until p broken triples triangles are satisfied. This is due to the fact that **min-triangle-breaking-node** is a special case of the **Partial Set Cover** problem and the greedy strategy guarantees an $H(p) - 1/2$ approximation solution, where $H(p)$ denotes the harmonic function $H(p) = 1 + 1/2 + \ldots + 1/p$. Thus, Algs. 1 and 2 are also $(H(p)-1/2)$-approximation algorithms for **min-triangle-breaking-node**.

C. ANALYSIS IN NETWORKS WITH POWER-LAW DEGREE DISTRIBUTION

We next show that the complexity of **DAK-n** can furthermore be reduced to $O(m^{3/2})$ in networks with power-law degree distributions, which are commonly exhibited in many real world complex systems such as the Internet, social, and biological networks [48], [49]. Conceptually, power-law degree distributed networks have the fraction of nodes with degree k (k connections to other nodes) is $\lfloor \frac{e^\alpha}{k^\gamma} \rfloor$, where e^α is the normalization factor as in the $P(\alpha, \gamma)$ model [50]. Practical networks usually have $2 < \alpha < 3$. In this work, we deduce the maximum degree in a $P(\alpha, \gamma)$ network to $e^{\frac{\alpha}{\gamma}}$ because for $k > e^{\frac{\alpha}{\gamma}}$, the number of edges will be less than 1. We show that in power-law degree distributed networks, the overall time complexity is $O(m^{3/2})$ which implies that **DAK-n** is as fast as the state-of-the-art algorithms for counting/listing triangles with no additional costs (Theorem 6). This also realizes the scalability of **DAK-n** in large networks.

Theorem 6: The complexity of **DAK-n** algorithm is $O(m^{\frac{3}{2}})$ on power-law degree distributed networks.

Proof: In a power-law degree distributed network, the numbers of vertices and edges are computed as follows,

$$n = \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} \frac{e^\alpha}{k^\gamma} \approx \begin{cases} \zeta(\gamma)e^\alpha & \text{if } \gamma > 1 \\ \alpha e^\alpha & \text{if } \gamma = 1, \\ \dfrac{e^{\frac{\alpha}{\gamma}}}{1-\gamma} & \text{if } \gamma < 1 \end{cases} \tag{5}$$

$$m = \frac{1}{2}\sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k\frac{e^\alpha}{k^\gamma} \approx \begin{cases} \dfrac{1}{2}\zeta(\gamma-1)e^\alpha & \text{if } \gamma > 2 \\ \dfrac{1}{4}\alpha e^\alpha & \text{if } \gamma = 2, \\ \dfrac{1}{2}\dfrac{e^{\frac{2\alpha}{\gamma}}}{2-\gamma} & \text{if } \gamma < 2 \end{cases} \tag{6}$$

where $\zeta(\gamma) = \sum_{i=1}^{\infty} \frac{1}{i^\gamma}$ is the Riemann Zeta function [50], [51] which converges absolutely for $\gamma > 1$ and diverges for all $\gamma \leq 1$. For the sake of simplicity, we will simply use real number instead of rounding down to integers.

The error terms can be easily bounded and are negligible in our proof.

Since Phase 1 of Alg. 2 is $O(m^{\frac{3}{2}})$ for counting triangles, we will analyze Phase 2 of Alg. 2 and show its overall complexity $O(m^{\frac{3}{2}})$. To this end, we first find the workload C_i at each round i in phase 2, sum them all up and utilize the power-law property to obtain the final result. In particular,

$$C_i = \sum_{v \in N(u_{max})} d_v.$$

The worst case of the second phase happens when $k = n$ which means that the algorithm has to select all nodes in decreasing order of triangle-breaking gains into the solution set S. That leads to the overall complexity of,

$$C = \sum_{i=1}^{n} C_i = \sum_{u \in V} \sum_{v \in N(u)} d_v = \sum_{u \in V} d_u^2. \quad (7)$$

We apply the power-law property on the number of nodes with degree k being $\frac{e^\alpha}{k^\gamma}$ and the maximum degree is $e^{\frac{\alpha}{\gamma}}$ on the above equation which yields

$$C = \sum_{u \in V} d_u^2 = \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k^2 \frac{e^\alpha}{k^\gamma} = e^\alpha \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k^{2-\gamma}. \quad (8)$$

We consider two cases:

Case 1 ($\gamma \geq 2$): This implies $k^{2-\gamma} \geq 1$. Eq. 8 becomes,

$$C = e^\alpha \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k^{2-\gamma} \leq e^\alpha \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} 1 = e^\alpha e^{\frac{\alpha}{\gamma}} = e^{\alpha + \frac{\alpha}{\gamma}}$$
$$\leq e^{\alpha + \frac{\alpha}{2}} = \left(e^\alpha\right)^{\frac{3}{2}}. \quad (9)$$

Combining Eq. 9 with the number of edges in power-law degree networks in Eq. 6, we obtain,

$$C \leq \left(e^\alpha\right)^{\frac{3}{2}} = \mathbf{c1} \cdot m^{\frac{3}{2}}, \quad (10)$$

where **c1** is a constant that satisfies,

$$\mathbf{c1} \approx \begin{cases} (\frac{1}{\frac{1}{2}\zeta(\gamma-1)})^{3/2} & \text{if } \gamma > 2. \\ (4/\alpha)^{3/2} & \text{if } \gamma = 2. \end{cases}$$

Note that $\gamma > 2$ infers $\zeta(\gamma - 1)$ converges and **c1** is a finite constant.

Thus, in this case, phase 2 has time complexity of $O(m^{\frac{3}{2}})$.

Case 2 ($\gamma < 2$): In this case, Eq. 8 is equivalent to,

$$C = e^\alpha \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k^{2-\gamma} = e^\alpha (e^{\frac{\alpha}{\gamma}})^{2-\gamma} \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} \frac{k^{2-\gamma}}{(e^{\frac{\alpha}{\gamma}})^{2-\gamma}}$$
$$\leq e^\alpha e^{\frac{2\alpha}{\gamma} - \alpha} \int_{t=0}^{1} t^{2-\gamma} \mathrm{d}t = e^{\frac{2\alpha}{\gamma}} \frac{1}{3-\gamma} = \mathbf{c2} \times m, \quad (11)$$

where

$$\mathbf{c2} \approx \frac{2(2-\gamma)}{3-\gamma},$$

is a finite constant since $\gamma < 2$. This yields the time complexity $O(m)$ for Phase 2. Finally, we conclude that the overall time complexity of $O(m^{\frac{3}{2}})$ in both cases.

VI. ALGORITHM FOR k-TRIANGLE-BREAKING-EDGE

In a similar vein to k-**triangle-breaking-node** and **min-triangle-breaking-node**, the edge removal variants expose similar attributes and thus the greedy algorithm can be directly applied with near-optimal guarantee. We present **DAK-e** for finding triangle-breaking edges in Alg. 3. On general networks, **DAK-e** performs faster than its node-version, **DAK-n** due to its complexity of $O(m^{3/2} + kn)$.

Algorithm 3 Discounting Algorithm for k-**Triangle-Breaking-Edge** (**DAK-e**)

Phase 1:

Renumber nodes so that $u < v$ implies $d(u) \leq d(v)$.
$F \leftarrow \emptyset$;
for each $(u, v) \in E$ **do** $tr(u, v) \leftarrow 0$;
for $u \leftarrow n$ to 1 **do**
    for each $v \in N(u)$ with $v < u$ **do**
        for each $w \in A(u) \cap A(v)$ **do**
            Increase $tr(u, v)$, $tr(v, w)$ and $tr(u, w)$ by one;
        Add u to $A(v)$;

Phase 2:

$Q \leftarrow$ Max-Priority-Queue(T)
For i = 1 **to** k
    $e_{max} \leftarrow Q.pop()$;
    Remove e_{max} from G and add e_{max} to F;
    Let $(u', v') = e_{max}$;
    for each $w \in N(u') \cap N(v')$ **do**
        Decrease $tr(w, u')$ and $tr(w, v')$ by one;
        $Q.update((w, u'), T)$;
        $Q.update((w, v'), T)$;
return F

In its execution **DAK-e** maintains, for each edge, the number of triangles incident on that edge and updates the measure efficiently when removing nodes from G. After removing an edge (u', v') we only need to consider only $|N(u') \cap N(v')|$ updates to discount the triangles incident on (u', v') from the corresponding edges. Thus the overall complexity in each iteration relies on finding the edge that breaks the maximum number of triangles. We obtain the following approximation guarantee for the edge-removal problem which is similar to the node version.

Theorem 7: **DAK-e** is an $(1 - 1/e)$-approximation algorithm for k-**triangle-breaking-edge** with complexity $O(m^{3/2} + kn)$.

On power-law degree distributed networks, by an argument similar to **DAK-n**, we can show that the overall complexity of **DAK-e** is $O(m^{\frac{3}{2}})$ which is also equal to counting/listing triangles in the networks as concluded in the following theorem.

Theorem 8: On power-law degree distributed networks, the complexity of DAK-e algorithm is $O(m^{\frac{3}{2}})$.

Note: An approach adapted algorithm from Alg. 3 can be devised for solving min-triangle-breaking-edge and returns a $(H(p) - 1/2)$-approximate edge set since min-triangle-breaking-edge is also a special case of Partial Set Cover problem.

A. INPUT-DEPENDENT APPROXIMATION GUARANTEES

The $(1 - 1/e)$-approximation factor, termed *fixed worst-case bound*, achieved by our algorithms provides a general lower-bound on the solution quality of the selected set S. This factor is known in advance even prior to the execution of the methods. Nevertheless, we can derive a better approximation bound of the solution quality, namely the *input-dependent bound*, depending on the problem instance and even the particular run of the algorithms. Inspired by the work in [20] on the Influence Maximization problem, we can apply a similar bounding technique (named online-bound) to obtain a real input-dependent bound on the solution quality in both the naive greedy and our DAK-n and DAK-e algorithms. The input-dependent bound for DAK-n is stated as follows,

Theorem 9 (DAK-n Input-Dependent Bound): For a set of selected nodes $S \subset V$ and each node $u \in V$, let $\Delta_S(u) = T(S \cup u) - T(S)$ be the marginal gain of u when u is included in S. Let $u_1, u_2, \ldots, u_{n-k}$ be the sequence of the remaining nodes (not in S) sorted in decreasing order of $\Delta_S(u)$, then

$$OPT_k^n \leq T(S) + \sum_{i=1}^{k} \Delta_S(u_i), \tag{12}$$

where $OPT_k^n = \max_{S' \subset V, |S'|=k} T(S')$ is the triangles broken by the optimal solution with k nodes.

By selecting the top k nodes with the highest marginal gains into the returned solution S of DAK-n, we obtain an upper-bound on the optimal solution. Then by dividing the number of triangles broken by S with that upper-bound, we have an input-dependent guarantee on S,

$$\mathcal{OB}_n(S) = \frac{T(S)}{T(S) + \sum_{i=1}^{k} \Delta_S(u_i)} \geq \frac{T(S)}{OPT_k^n}. \tag{13}$$

Similarly, the input-dependent for solution F of the DAK-e is computed by the following equation,

$$\mathcal{OB}_e(F) = \frac{T(F)}{T(F) + \sum_{i=1}^{k} \Delta_F(e_i)} \geq \frac{T(F)}{OPT_k^e}, \tag{14}$$

where $e_1, \ldots, e_k$ are the top k edges with the highest marginal gain of broken triangles with respect to F and OPT_k^e is the triangles broken by the optimal edge set with k edges.

VII. EXPERIMENTAL EVALUATION

In this section, we evaluate the quality and performance of our proposed methods DAK-n and DAK-e. We show, through empirical results, two important features of our approaches: *performance* and *scalability* that are desired for any practical techniques. We compare and contrast ours with GreedyAll,[1] the state-of-the-art method suggested in [21], and approaches based on centralites, i.e., Max-degree, Pagerank and Randomization. Betweeness centrality was not included in the experiments due to its time consumption in large networks. On k-triangle-breaking-node and k-triangle-breaking-edge problems, results indicate that our methods vastly outperform GreedyAll up to 100× in terms of time consumption while achieving the same level of solution quality. The baseline methods based on centrality and randomization are slightly faster but the qualities are much worst. We also spend a good portion to study the networks under node and edge removal attacks using the min-triangle-breaking-node and min-triangle-breaking-edge problems.

A. EXPERIMENTAL SETTINGS

1) DATASETS

To make our experiments extensive, we select a set of six real-world traces from various domains with sizes ranging from thousand to million scales. The summary of those networks are provided in Table. 3.

Specifically, our dataset includes both physical (connected by physical links) and virtual (e.g., friendship, communication) networks. In the first category: Gnutella4 is a snapshot of the Gnutella peer-to-peer file sharing network on

[1] Work in [21] also proposed **Approx** which used FM-sketch to approximate the triangle-breaking gain. This approach imposes the same time complexity with GreedyAll.

(*) http://snap.stanford.edu/data/index.html;

(y) http://socialcomputing.asu.edu/pages/datasets

TABLE 3. Real-World Networks for Experimentation.

Dataset	Type	#Nodes	#Edges	Avg. degree
Gnutella4	Peer-to-peer network(*)	10.9K	40K	3.7
Flickr	Photo sharing network(†)	80.5K	11.8M	138.8
Google	Web graph(*)	876K	5.1 M	5.83
Skitter	Internet Topology(*)	1.7M	11.1M	6.53
Wiki-Talk	Wikipedia Communication(*)	2.4M	5M	2.1
Orkut	Online Social Network(*)	3M	117M	78

August 4th 2002 in which nodes represent hosts in the Gnutella network topology and edges represent connections between the hosts; Skitter is the Internet topology graph captured by tracerouting in 2005. In the second category: Flickr is a contact network crawled from the photo sharing Flickr website where nodes are users and edges are friendship connections between users; Google is the dataset of web-pages and hyperlinks between the webs released by Google company in 2002; Wiki-Talk contains the set of users in the Wikipedia website and edit relationship (who edits take pages of whom) and Orkut is an online social networks with users as nodes and friendships as connections.

2) PERFORMANCE AND SCALABILITY MEASURES

(Performance) For a fair comparison between different methods, we count the number of triangles broken by the set of nodes/edges returned by the algorithms as the quality measure.

(Scalability) In terms of scalability, we record the running time consumed by each algorithm. For the **min-triangle-breaking-node** and **min-triangle-breaking-edge** problem, we only measure the running time of **DAK-n** and **DAK-e**. The input-dependent bound of our algorithms is also illustrated in the last experiments.

3) IMPLEMENTATION AND TESTING ENVIRONMENT

We implemented our algorithms **DAK-n** and **DAK-e** in C++ programming language with GCC 4.8 C++11 compiler. We also implemented the GreedyAll [21] algorithm following closely the provided description and pseudo-code. All the experiments are run on a Linux environment with 2.2Ghz Xeon 8 core processor and 100GB of RAM. In each execution, only a single core is assigned for each method.

B. PERFORMANCE EVALUATION

The performance, i.e., the solution quality, measured by the number of triangles broken by node and edge removals is illustrated in Figures. 1 and 2 for node and edge variants, respectively. In this evaluation, the higher the number of triangles disconnected by the removal of sets of k-nodes/edges the better. As depicted from these figures, **DAK-n**, **DAK-e** and GreedyAll consistently have the best performance on all the social traces compared to the others. This indicates node and edge sets selected by those algorithms are crucial in maintaining the network's clustering and strong connectivity. In terms of social networks, the nodes and edges identified can be considered as important or influential users/relationships that are key to the network's function. Pagerank and Max-degree achieve very good solution quality on certain datasets, e.g., Google and Wiki-Talk, but fall far behind **DAK-n**, **DAK-e** and GreedyAll on the other tests. The quality of Random strategy, as expected, falls below and is inconsistent compared to the others. In summary, empirical results from multiple real-world data confirm the performance provided by our suggested algorithms. Figures. 1 and 2 also display the typical trend of monotone and submodular functions as they exhibit the diminishing return property. For the first few selections, the marginal gain (in terms of the number of broken triangles)

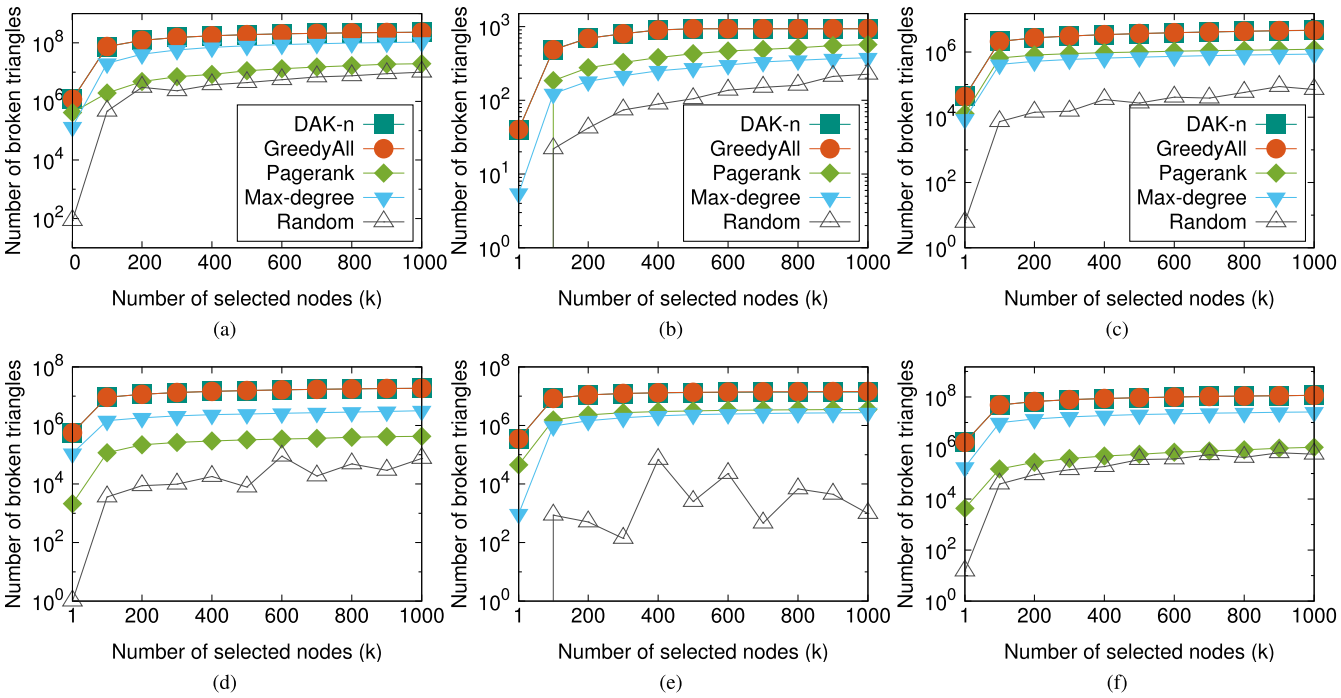


FIGURE 1. Number of broken triangles by node removal algorithms (higher value is better). (a) Flickr. (b) Gnutella. (c) Google. (d) Skitter. (e) Wiki-Talk. (f) Orkut.

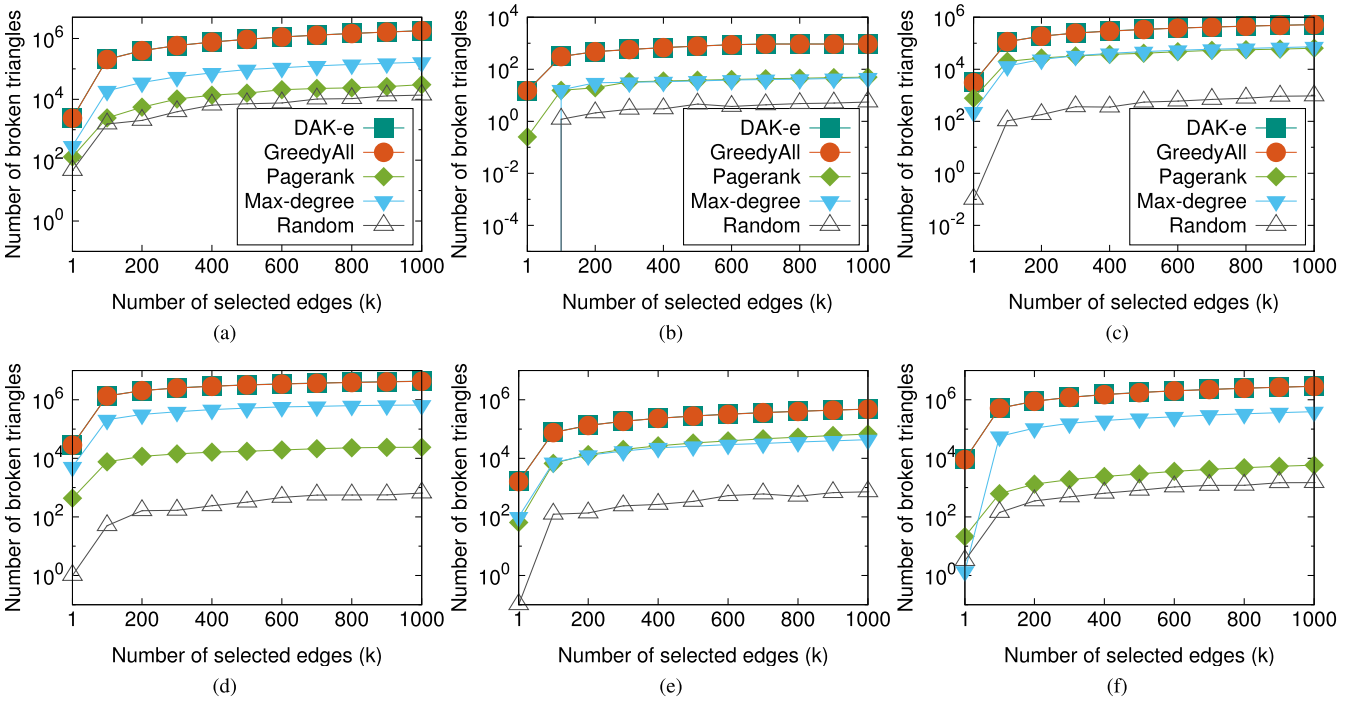


FIGURE 2. Number of broken triangles broken by edge removal algorithms (higher value is better). (a) Flickr. (b) Gnutella. (c) Google. (d) Skitter. (e) Wiki-Talk. (f) Orkut.

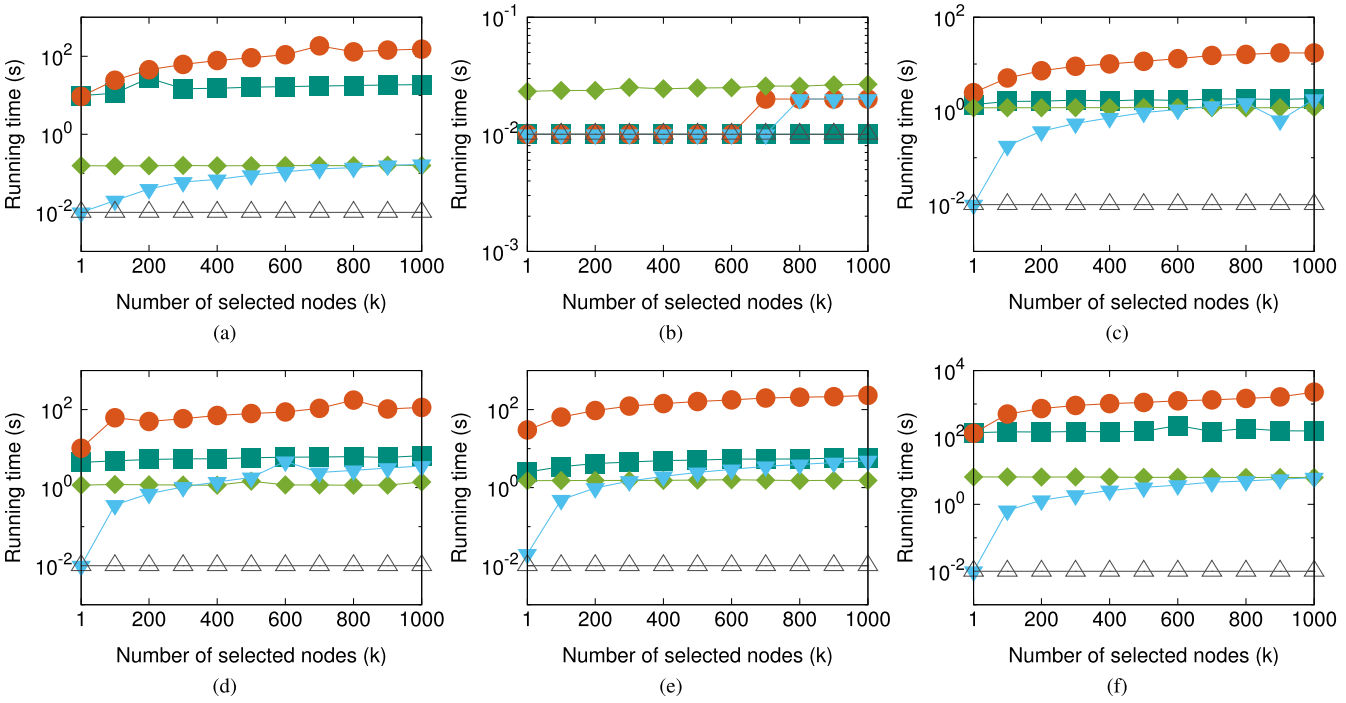


FIGURE 3. Running time of node removal algorithms (legends in Figure 1). (a) Flickr. (b) Gnutella. (c) Google. (d) Skitter. (e) Wiki-Talk. (f) Orkut.

is significant yet the later rounds provide smaller marginal gain, and the gain tends to saturate quickly.

C. SCALABILITY EVALUATION

Figures. 3 and 4 report the time consumption (in seconds) of testing algorithms in experiments. These figures display three groups of methods with different magnitudes: (Group 1) GreedyAll with most time consumption (up to 100× times higher than the second group), (Group 2) DAK-n, DAK-e, Pagerank and Max-degree algorithms, and (Group 3) Random method which returns almost instantly k random nodes/edges due to it simple nature. We observe

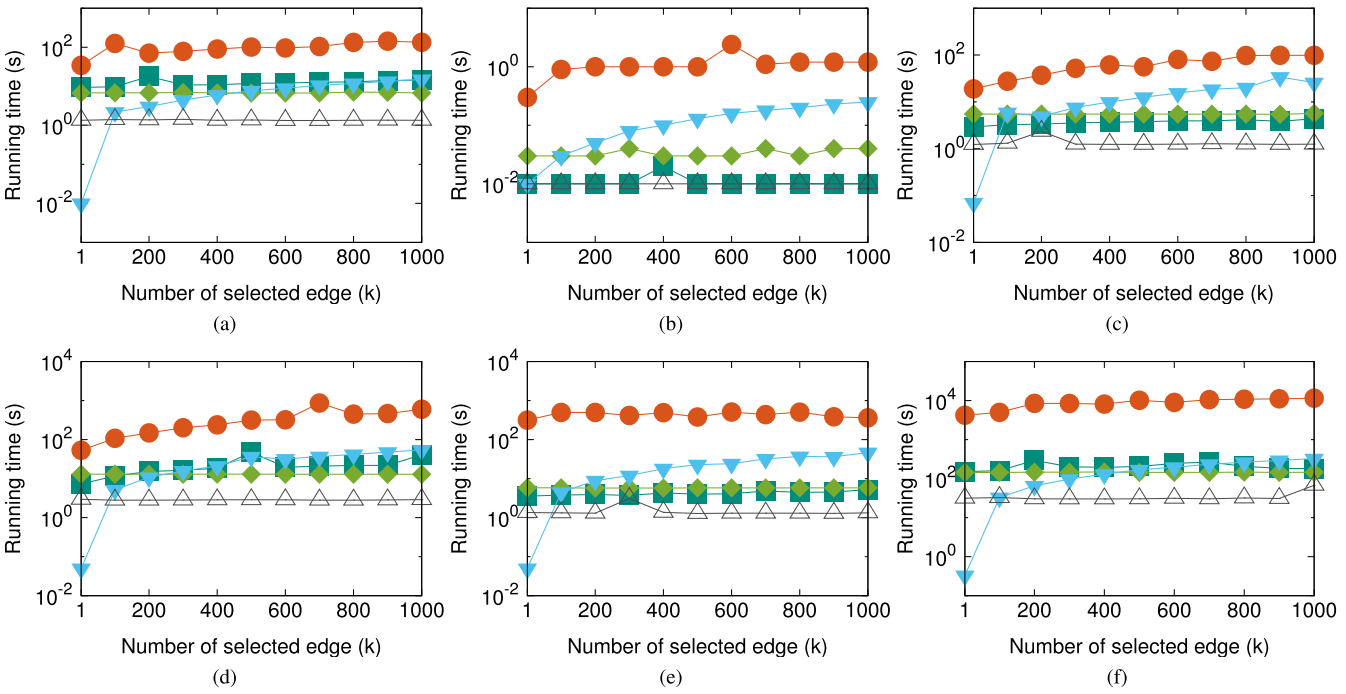


FIGURE 4. Running time of edge removal algorithms (legends in Figure 1). (a) Flickr. (b) Gnutella. (c) Google. (d) Skitter. (e) Wiki-Talk. (f) Orkut.

that DAK-n and DAK-e require a very comparable amount of time to Pagerank and Max-degree methods, the two canonical centralities and very fast to compute. Better yet, DAK-n and DAK-e produce much better solution quality than Pagerank and Max-degree while are very comparable in terms of scalability.

These extensive experiments illustrate that our proposed DAK-n and DAK-e algorithms is highly competitive to the current best GreedyAll method performance meanwhile is much better in terms of scalability. As shown in the previous experiments, only GreedyAll has similarly highest level of solution quality as DAK-n and DAK-e; however, our running time results show that GreedyAll is up to 20 slower than DAK-n on the node removal problem and 100 times slower than DAK-e on the edge removal variants.

D. INPUT-DEPENDENT BOUND TESTING

Finally, we perform experiments on the input-dependent bounding technique embedded in DAK-n and DAK-e algorithms. Theoretically, the solutions returned by DAK-n and DAK-e are guaranteed to be at least $(1 - 1/e) \approx 0.63$ on any input instance. In practice, we can even obtain better guarantees depending on the problem instance and the execution itself. Our input-dependent bounding strategy is one way of finding such instance- and execution-dependent guarantees. Table 4 presents the input-dependent bounds provided by DAK-n algorithm for node removal problem. Values in this table express the input-dependent bounds and the closer to 1 the better. We can observe that these input-dependent bounds are substantially better than the theoretical guarantee $1 - 1/e \approx 0.63$. For example, with $k = 400$ on Wiki-Talk dataset, DAK-n is guaranteed to output a solution whose quality is at least 95% of the optimal one. For Gnutella network, with $k \geq 600$, DAK-n indeed finds the optimal solution, indicating that all the triangles in the network have been disconnected. We can also observe that the bound gets tighter when k increases. This implies our suggested algorithms

TABLE 4. Input-Dependent Bounds Provided by DAK-n (Closer to 1 Is Better).

Data	$k = 200$	$k = 400$	$k = 600$	$k = 800$	$k = 1000$
Flickr	0.65	0.74	0.81	0.85	0.88
Gnutella	0.77	0.90	1	1	1
Google	0.78	0.78	0.78	0.79	0.79
Skitter	0.77	0.80	0.82	0.84	0.85
Wiki-Talk	0.84	0.95	0.97	0.99	0.99
Orkut	0.75	0.79	0.81	0.81	0.82

closely approach the optimal solutions (which are generally very hard to find out due to their NP-hardness) as more nodes are allowed in the budget k. The reason behind that is due to the nature of our bounding technique: larger k means more triangles are broken and the gain of the next k nodes becomes smaller and approximation ratio approaches 1.

In summary, we observe that our proposed algorithms (1) perform much better in practice with less time consumption in comparison with other methods, and (2) obtain much tighter (and sometimes the best) approximation guarantees than the theoretical bounds as the budget k increase. These features indicate the applicability of our approaches for real-world social network data.

VIII. CONCLUSION

We investigate critical nodes and links whose failures will severely damage most triangles in the network, changing the network's organization and (possibly) leading to the unpredictable dissolving of the network. We formulate this vulnerability analysis as optimization problems, and provide proofs of their NP-Completeness. We propose two algorithms DAK-n and DAK-e with notable performance and scalability. Both DAK-n and DAK-e obtain best approximation guarantees: 19/27-approximation for k-triangle-breaking-node and k-triangle-breaking-edge as well as 3-approximation for min-triangle-breaking-node and min-triangle-breaking-edge, and are scalable for network with millions nodes and edges. Those features lend our approaches nicely into the analysis of various large-scale real-world problems. In the future, we aim to bridge the gaps between theory and practice to design more scalable approaches with better approximation guarantees.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 65, no. 5, p. 056109, May 2002.

[2] N. P. Nguyen, T. N. Dinh, D. T. Nguyen, and M. T. Thai, "Overlapping community structures and their detection on social networks," in *Proc. IEEE SocialCom*, Oct. 2011, pp. 35–40.

[3] N. P. Nguyen, Y. Xuan, and M. T. Thai, "A novel method for worm containment on dynamic social networks," in *Proc. IEEE MILCOM*, Oct. 2010, pp. 2180–2185.

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[5] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, May 2007, Art. no. 5. [Online]. Available: http://doi.acm.org/10.1145/1232722.1232727

[6] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.

[7] K. J. Barclay, C. Edling, and J. Rydgren, "Peer clustering of exercise and eating behaviours among young adults in Sweden: A cross-sectional study of egocentric network data," *BMC Public Health*, vol. 13, no. 1, p. 784, 2013.

[8] L. Lü, D.-B. Chen, and T. Zhou, "The small world yields the most effective information spreading," *New J. Phys.*, vol. 13, no. 12, p. 123005, 2011.

[9] N. Malik and P. J. Mucha, "Role of social environment and social clustering in spread of opinions in coevolving networks," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 23, no. 4, p. 043123, 2013.

[10] D. Centola, "An experimental study of homophily in the adoption of health behavior," *Science*, vol. 334, no. 6060, pp. 1269–1272, 2011.

[11] D. Boyd, "Friendster lost steam. Is myspace just a fad?" *Apophenia Blog*, vol. 1, no. 1, p. 1, 2016.

[12] H. Travis. (2013). *Cyberspace Law: Censorship Regulation Internet*, Routledge. [Online]. Available: https://books.google.com/books?id=-WOAAAAAQBAJ

[13] J. Ponton, P. Wei, and D. Sun, "Weighted clustering coefficient maximization for air transportation networks," in *Proc. Eur. Control Conf. (ECC)*, Jul. 2013, pp. 866–871.

[14] M. A. Alim, X. Li, N. Nguyen, M. Thai, and A. Helal, "Structural vulnerability assessment of community-based routing in opportunistic networks," *IEEE Trans. Mobile Comput.*, vol. 15, no. 12, pp. 3156–3170, Dec. 2016.

[15] T. N. Dinh, Y. Xuan, M. T. Thai, P. M. Pardalos, and T. Znati, "On new approaches of assessing network vulnerability: Hardness and approximation," *IEEE/ACM Trans. Netw.*, vol. 20, no. 2, pp. 609–619, Apr. 2012.

[16] H. Chan, H. Tong, and L. Akoglu, *Make It or Break It: Manipulating Robustness in Large Networks*. Philadelphia, PA, USA: SIAM, 2014, ch. 37, pages 325Ű333.

[17] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, Jul. 2000.

[18] S. Allesina and M. Pascual, "Googling food Webs: Can an eigenvector measure species' importance for coextinctions?" *PLoS Comput. Biol.*, vol. 5, no. 9, p. e1000494, 2009.

[19] T. Dinh and R. Tiwari, *Breaking Bad: Finding Triangle-Breaking Points in Large Networks*. Cham, Switzerland: Springer, 2015, pp. 285–295.

[20] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. ACM KDD*, New York, NY, USA, Aug. 2007, pp. 420–429.

[21] R.-H. Li and J. X. Yu, "Triangle minimization in large networks," *Knowl. Inf. Syst.*, vol. 45, no. 3, pp. 617–643, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1007/s10115-014-0800-9

[22] T. H. Grubesic, T. C. Matisziw, A. T. Murray, and D. Snediker, "Comparative approaches for assessing network vulnerability," *Int. Regional Sci. Rev.*, vol. 31, no. 1, pp. 88–112, 2008.

[23] A. Murray, T. Matisziw, and T. Grubesic, "Multimethodological approaches to network vulnerability analysis," *Growth Change*, vol. 39, no. 4, pp. 573–592, 2008.

[24] S. Neumayer, G. Zussman, R. Cohen, and E. Modiano, "Assessing the vulnerability of the fiber infrastructure to disasters," *IEEE/ACM Trans. Netw.*, vol. 19, no. 6, pp. 1610–1623, Dec. 2011.

[25] T. N. Dinh and M. T. Thai, "Precise structural vulnerability assessment via mathematical programming," in *Proc. IEEE MILCOM*, Nov. 2011, pp. 1351–1356.

[26] T. N. Dinh and M. T. Thai, "Network under joint node and link attacks: Vulnerability assessment methods and analysis," *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 1001–1011, Jun. 2015.

[27] H. Frank and I. Frisch, "Analysis and design of survivable networks," *IEEE Trans. Commun. Technol.*, vol. 18, no. 5, pp. 501–519, Oct. 1970.

[28] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Math. J.*, vol. 23, no. 2, pp. 298–305, 1973.

[29] T. P. Peixoto and S. Bornholdt, "Evolution of robust network topologies: Emergence of central backbones," *Phys. Rev. Lett.*, vol. 109, no. 11, pp. 118703–118708, 2012.

[30] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and fragility: Percolation on random graphs," *Phys. Rev. Lett.*, vol. 85, no. 25, pp. 5468–5471, Dec. 2000.

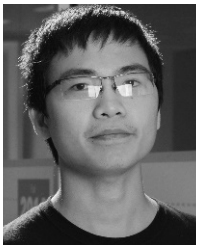
[31] T. N. Dinh, D. T. Nguyen, and M. T. Thai, "Cheap, easy, and massively effective viral marketing in social networks: Truth or fiction?" in *Proc. 23rd ACM Conf. Hypertext Social Media*, Jun. 2012, pp. 165–174.

[32] T. N. Dinh, H. Zhang, D. T. Nguyen, and M. T. Thai, "Cost-effective viral marketing for time-critical campaigns in large-scale social networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 2001–2011, Dec. 2014.

[33] T. N. Dinh, N. P. Nguyen, and M. T. Thai, "An adaptive approximation algorithm for community detection in dynamic scale-free networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 55–59.

[34] 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. IEEE 13th Int. Conf. Data Mining (ICDM)*, Dec. 2013, pp. 567–576.

[35] T. N. Dinh and M. T. Thai, "Toward optimal community detection: From trees to general weighted networks," *Internet Math.*, vol. 11, no. 3, pp. 181–200, 2015.
[36] N. P. Nguyen, M. A. Alim, Y. Shen, and M. T. Thai, "Assessing network vulnerability in a community structure point of view," in *Proc. IEEE/ACM ASONAM*, Aug. 2013, pp. 231–235.
[37] M. A. Alim, N. P. Nguyen, T. N. Dinh, and M. T. Thai, "Structural vulnerability analysis of overlapping communities in complex networks," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*. New York, NY, USA, Aug. 2014, pp. 231–235.
[38] T. Schank and D. Wagner, "Finding, counting and listing all triangles in large graphs, an experimental study," in *Proc. 4th Int. Conf. Experim. Efficient Algorithms*, 2005, pp. 606–609.
[39] N. Alon, R. Yuster, and U. Zwick, "Finding and counting given length cycles," *Algorithmica*, vol. 17, no. 3, pp. 209–223, 1997. [Online]. Available: http://dx.doi.org/10.1007/BF02523189
[40] S. Suri and S. Vassilvitskii, "Counting triangles and the curse of the last reducer," in *Proc. 20th Int. Conf. World Wide Web*, Mar./Apr. 2011, pp. 607–614.
[41] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, "Reductions in streaming algorithms, with an application to counting triangles in graphs," in *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2002, pp. 623–632.
[42] L. S. Buriol, G. Frahling, S. Leonardi, A. Marchetti-Spaccamela, and C. Sohler, "Counting triangles in data streams," in *Proc. 25th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, Jun. 2006, pp. 253–262.
[43] H. Jowhari and M. Ghodsi, "New streaming algorithms for counting triangles in graphs," in *Computing and Combinatorics*. Berlin, Germany: Springer, 2005, pp. 710–716.
[44] R. Gandhi, S. Khuller, and A. Srinivasan, "Approximation algorithms for partial covering problems," *J. Algorithms*, vol. 53, no. 1, pp. 55–84, Oct. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196677404000689
[45] A. A. Ageev and M. I. Sviridenko, "Pipage rounding: A new method of constructing algorithms with proven performance guarantee," *J. Combinat. Optim.*, vol. 8, no. 3, pp. 307–328, Sep. 2004.
[46] V. V. Vazirani, *Approximation Algorithms*. Berlin, Germany: Springer, 2001. [Online]. Available: http://books.google.com/books?id=EILqAmzKgYIC
[47] M. Yannakakis, "Edge-deletion problems," *SIAM J. Comput.*, vol. 10, no. 2, pp. 297–309, 1981.
[48] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the world-wide Web," *Phys. A, Statist. Mech. Appl.*, vol. 281, nos. 1–4, pp. 69–77, Jun. 2000.
[49] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Phys. A, Statist. Mech. Appl.*, vol. 311, nos. 3–4, pp. 590–614, Aug. 2002.
[50] W. Aiello, F. Chung, and L. Lu, "Random evolution in massive graphs," in *Handbook of Massive Datasets*. Norwell, MA, USA: Kluwer, 2001, pp. 510–519.
[51] T. N. Dinh and M. T. Thai, "Community detection in scale-free networks: Approximation algorithms for maximizing modularity," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 6, pp. 997–1006, Jun. 2013.



HUNG T. NGUYEN received the B.S. degree in information technology from Vietnam National University, Hanoi, Vietnam, in 2014. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Virginia Commonwealth University, under the supervision of Dr. T. N. Dinh. His research focuses on designing efficient approximation algorithms with optimality guarantees for problems in influence diffusion, such as influence maximization, infection source identification, on billion-scale networks under dynamic stochastic settings, and finding community structures in heterogeneous multiplex networks.

NAM P. NGUYEN received the B.S. degree from Vietnam National University, in 2007, and the M.S. degree from Ohio University, in 2009, and the Ph.D. degree in computer engineering from the University of Florida in 2013. He is currently an Assistant Professor with the Computer and Information Sciences Department, Towson University. His research interests focus on complex network structure and analysis, big data and social aware information mining, cyber security, and mobile computing. He serves as a TPC member for several conferences including the INFOCOM, the IJCAI, and the WASA. He was the TPC chair of the CSoNet'15 and the BCD'16. He also serves on the Editorial Board of Springer's *Computational Social Network* Journal. He was a recipient of The Jess and Mildred Fisher Endowed Professor of Computer Science Award (2016–2019), Towson University.

TAM VU (GS'12–M'15) is currently an Assistant Professor with the Department of Computer Science, University of Colorado, USA. His research interests include designs and implements novel and practical cyber-physical systems to make physiological sensing, including breathing volume measurement, brainwave signal monitoring, muscle movement recording, sleep quality monitoring, and less intrusive at lower cost. He was a recipient of four Best Paper Awards at the ACM SenSys 2016, the MobiCom 2016 S3, the MobiCom 2012, and the MobiCom 2011. He received the Google Faculty Research Award in 2014; the Creative Research Fellowship of UC Denver in 2015, the wide press coverage including the Denver Post, the CNN TV, the Fox News Channel, the National Public Radio, The NY Times, The Wall Street Journal, MIT TechReview, and Yahoo! News.

HUAN X. HOANG received the Ph.D. degree the Faculty of Mathematics, Hanoi University, with a focus on separable calibration and globally minimal currents. He was a Lecturer with the Faculty of Mathematics, Hanoi University, from 1980 to 1995. In 1995, he joined the Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam, where he is currently an Associate Professor. He has authored three books and over 50 papers, covering a large spectrum of topics including optimization techniques, evolutionary computations, economic mathematics, machine learning, and bioinformatics.

THANG N. DINH (S'11–M'14) received the Ph.D. degree in computer engineering from the University of Florida, in 2013. He is currently an Assistant Professor with the Department of Computer Science, Virginia Commonwealth University. His research focuses on security and optimization challenges in complex systems, especially social networks, wireless, and cyber-physical systems. He serves as the PC Chair of the COCOON'16 and the CSoNet'14 and the TPC of several conferences including the IEEE INFOCOM, the ICC, the GLOBECOM, and the SOCIALCOM. He is also an Associate Editor of the *Computational Social Networks* journal and a Guest-Editor of the *Journal of Combinatorial Optimization*.