Internet Mathematics, 12:145–204, 2016

ISSN: 1542-7951 print/1944-9488 online
DOI: 10.1080/15427951.2015.1098755

A THEORY OF NETWORK SECURITY: PRINCIPLES OF NATURAL SELECTION AND COMBINATORICS

Angsheng Li[1] and Yicheng Pan[1,2]
[1] *State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, P. R. China*
[2] *State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, P. R. China*

Abstract *We propose the definition of* security *of networks against the cascading failure models of deliberate attacks. We propose a model of networks by the natural selection of homophyly/kinship, randomness, and preferential attachment, referred to as the* security model. *We show that the networks generated by the security model are provably secure against any attacks of sizes* poly$(\log n)$ *under the cascading failure models, for which the principles of natural selection and the combinatorial principles of the networks of the security model, including a power law, a self-organizing principle, a small-diameter property, a local navigation law, a degree-priority principle, an inclusion–exclusion principle, and an infection priority tree principle, etc., are the underlying principles. Furthermore, we show that the networks generated by the security model have an expander core. This property ensures that the networks of the security model satisfy the requirement of global communications in engineering. Based on our theory, we propose a security protocol for computer networks. Our theory demonstrates that security of networks can be achieved by a merging of natural selection and combinatorial principles, and that both natural selection principles, and combinatorial principles are essential to security of networks.*

1. INTRODUCTION

Real-world networks satisfy a number of statistical properties, including the well-known power-law degree distributions, small-world property, high clustering coefficients, hierarchical properties, and various centralities [14]. A number of network models have been proposed to analytically understand these properties. In a long history of research of complex systems, it has been assumed that complex networks are random, for which the classic Erdös–Rényi (ER) [7, 15, 16] model serves as a generator. The ER model generates graphs with small-diameters, predicting that the small-diameter property may be shared by many networks. To explain the power-law degree distribution of some real-world networks, [3, 4] proposed a model by introducing *preferential attachment* as an explicit mechanism, referred to as (the PA) model. It was shown that PA networks do follow a power law.

That the WWW graph is rich in bipartite cliques was discovered by [23]. Based on this, [22] proposed the copying model by introducing a copying rule in random models to generate power-law graphs in which there are rich bipartite cliques, interpreted as

Address correspondence to Angsheng Li, Institute of Software, Chinese Academy of Sciences, 4 Zhongguancun South 4th St., Haidian, 100190, Beijing, P. R. China. Email: angsheng@ios.ac.cn

Color versions of one or more figures in the article can be found online at www.tandfonline.com/uinm.

communities. The idea of copying edges was also used in other models such as the forest fire model [25], random walk model, nearest neighbor model [46], and random-surfer model [5]. Each of these models uses randomness and some local rules including the copying mechanism. The forest fire model generates dense graphs with power law, for which the graphs are claimed to have a community structure as a result of the copying operation. The random walk model, the nearest neighbor model, and the random-surfer model generate power-law graphs in which clustering coefficients are amplified due to the creation of more triangles by copying. These models demonstrate that power-law graphs may be generated by randomness together with some local rules rather than by PA. In addition, the networks generated by the models with certain local rules inspired by the copying actions could have interesting subgraphs, such as the connected union of k-cliques for $k \geq 3$ and bipartite cliques, etc. These features together with other properties have been analyzed in [46].

The small-world phenomenon of networks consists of two properties: the first is the small-diameter property, and the second is the clustering effect in the sense that two nodes are more likely to be adjacent if they share common neighbors. To capture the clustering effect, [48] proposed a simple model to add random edges to a grid graph or the like and [21] introduced the model of adding edges with endpoints chosen with probability inversely proportional to a power of the distances in the grid. Additionally, because randomness is the unique mechanism of the ER model that leads to the small-diameter property, we conjecture that randomness is the mechanism of small-diameter phenomenon for many networks. The conjecture indicates that the small-diameter property of networks is simply the result of randomness. This understanding regards randomness as a useful resource in networking systems, leading to the motivation to fully explore the role of randomness in networks.

A hierarchical model starting from a fixed initial module by iteratively and recursively copying the module was proposed by [41]. This idea was extended to a model by using Kronecker multiplication in [24]. Graphs generated from these models usually contain densely connected subgraphs recursively generated from a fixed module to be interpreted as communities. Each of these existing models explores specific features of real-world networks and generates graphs that are rich in interesting subgraphs.

The hymophyly/kinship hypothesis that homophyly is the extension of Darwinian selection of kinship, and that homophyly is the natural selection that controls the organization and evolution of real-world, high-level networks from random variation was proposed by [26], and the hypothesis was validated by experiments on some real-world networks. Further, [27] proposed the notion of structure entropy of graphs and a greedy algorithm minimizing the structure entropy of networks such that the algorithm identifies natural communities of networks both model generated and naturally evolving networks in society. The results in [26, 27] demonstrated that homophyly is the mechanism of natural community structures of networks, and that minimization of structure entropy or, equivalently, minimization of the nondeterminism of structures, is both the principle of self-organization of networks and the principle for detecting natural communities in networks.

The existing theoretical study on networks is largely devoted to the *static theory of networks* such as the mechanisms, statistical properties, and combinatorics of networks. In sharp contrast to the static research, there are few theoretical studies on the *dynamics of networks*. However, dynamics of networks include a number of fundamental challenges such as security and robustness of networks against attacks and the emergence of cooperation in evolutionary games in networks. These issues are too important to allow them to fail.

Network security has been a fundamental issue from the very beginning of network science. In the last few years, security of networks has become an urgent challenge in network applications.

The issues of network security differ for different types of attacks. Typical attacks include both *physical attack* of removal of nodes or edges and *cascading failure models of attacks*, which are similar to viruses spreading.

The first type is the physical attack of removal of nodes, for which the goal is to destroy the global connectivity of networks. In [1,11,37], it is shown that in the power-law networks of the PA model, the overall network connectivity measured by the sizes of the giant connected components and the diameters does not change significantly under random removal of a small fraction of nodes, but is vulnerable to removal of a small fraction of the high-degree nodes. This progress suggests an issue of security and robustness against physical attacks or random errors of removal of nodes or edges.

The second type of attack is the cascading failure model, see for instance [2,17,36, 43,47]. This model captures the behaviors of spreading of information, viruses on computer networks, news on the Internet, ideas on social networks, and of influence in economical networks, etc. There are different definitions of diffusions in networks in the literature. Here, we investigate the *threshold cascading failure model*, which was formulated in social studies and used in simulating the epidemic spread in networks [19]. In this model, the members have a binary decision and are influenced by their neighbors in the scenarios of rumor spreading, disease spreading, voting, advertising, etc. This model of cascading behavior has been studied in physics, sociology, biology, and economics [2,18,36,38,47]. In [40], it is shown that in the power-law networks of the PA model, even a weakly virulent virus can spread. This research leads to the issue of security against cascading failures.

The algorithmic aspect of the threshold cascading failure model on regular graphs of different patterns, particularly on cliques and trees, has been studied [6]. A $(1-\frac{1}{e})$-approximation algorithm, which, for a given k, finds a set S of size k such that the influence of S is at least a $(1-\frac{1}{e})$ fraction of that of the most influential k nodes, where the influence of a set in a graph is the set of the nodes that are infected by the set is given in [20].

It has been experimentally shown that the edges among different communities are more important than the high-degree nodes for influence spreading in social networks [50]; and [42] experimentally showed that immunization interventions that targeted the individuals bridging communities are more effective than those simply targeting highly connected individuals; this demonstrates that community structures strongly affect disease dynamics. This progress suggests that a strong community structure of a network allows an effective way to control disease-spreading by controlling the edges of nodes between different communities.

A strategy for network security and a security model of networks was proposed by [29] and numerically verified the thesis of defensive stratagem of the Art of War [44]. In the third of the 13 chapters of the *Art of War* [44] (published in 512 BC), by ancient Chinese Sun Wu, the Chinese military strategist (535–?, BC), the proposed thesis is that the best strategy is the attack that frustrates the enemy's strategy–the attack of rules of networks; that the second best strategy is the attack through diplomacy–the attack by cascading failure of the virus; that the third best strategy is the attack of the enemy's army–the physical attack of deleting nodes or edges; and that the last strategy is to attack cities–to destroy the structures of networks. Our theory in [29] and in the present article demonstrates that the thesis [44] is perfectly correct, which can be proved both numerically and mathematically.

In the present study, we establish the first *network dynamics* as a *security theory of networks*. Our theory is a *global theory of networks* resolving the issue of security of networks against deliberate attacks. We establish our theory by proving the experimental results discovered in [29]. We organize the article as follows. In Section 2, we define the infection and injury set of cascading failures and of physical attacks, respectively, and verify that cascading failure is more serious than physical attacks, and that physical attack could be a special case of cascading failure of attacks. In Section 3, we define the notion of security of networks. In Section 4, we introduce five representative existing models of networks and verified that none of the models generates secure networks. In Section 5, we introduce our security model of networks. In Section 6, we introduce the theorems of the security model. In Section 7, we prove the fundamental theorem of the networks generated by the security model. In Section 8, we prove some combinatorial principles of the networks of the security model, which are crucial to the proofs of our security theorems. In Section 9, we establish an infection priority tree principle. In Section 10, we prove our security theorems. In Section 11, we establish an expander core theorem of the networks generated by the security model. In Section 12, we propose a security protocol for computer networks on the basis of analyses of the combinatorial principles of the networks of the security model and the engineering networking of computer networks. At the end of the article, we summarize our conclusions and discuss some future directions.

2. INFECTION SET AND INJURY SET

In this section, we introduce the basic definitions for us to quantitatively analyze the security of networks.

We define the threshold cascading failure model as follows.

Definition 2.1. **(Infection set.)** Let $G = (V, E)$ be a network. Suppose that for each node $v \in V$, there is a threshold $\phi(v)$ associated with it. For an initial set $S \subset V$, the *infection set* of S in G is defined by stages as follows:

Stage 0. Each node $x \in S$ is called infected. Equivalently, we define a set I_0 such that $I_0 = S$.

For $i \geq 0$, suppose that I_i has been defined.

Stage $i + 1$. If there is a node $v \in V \setminus I_i$ such that $\phi(v)$ fraction of v's neighbors are in I_i, then enumerate all such nodes v's in I_{i+1}. Otherwise, then terminate with output I_i.

We call the output set I_i of the construction above the infection set of S in G, and denote it by $\inf_G(S)$.

The cascading failure models depend on the choices of thresholds $\phi(v)$ for all v. We consider two natural choices of the thresholds. The first is random threshold cascading, and the second is uniform threshold cascading.

Definition 2.2. **(Random threshold.)** We say that a cascading failure model is random if, for each node v, $\phi(v)$ is defined randomly and uniformly, that is, $\phi(v) = r/d$, where d is the degree of v in G, and r is chosen randomly and uniformly from $\{1, 2, \ldots, d\}$.

Definition 2.3. **(Uniform threshold.)** We say that a cascading failure model is uniform if, for each node v, $\phi(v) = \phi$ for some fixed number ϕ.

Remark. The security of a network certainly depends on a threshold function $\phi : V \rightarrow (0, 1]$. In the real-world networks, there are numerous choices for the threshold function ϕ. Our theory is to explore the role of network structures in the security of networks. For this purpose, we must assume the natural and most frequently appearing threshold functions are ϕ's. We argue that the random threshold and the uniform threshold functions have been already very general and representative for a theory of network security. Recall that, for a node v, the threshold $\phi(v)$ is the resistance of v from being infected from all its neighbors. From this understanding, the random threshold function assumes that the thresholds for all the nodes are randomly chosen. In computer networks, users usually install certain softwares to protect their computers. However, the softwares chosen by different users vary a great deal because the roles of the softwares are very much different. Our random threshold function captures the type of cascading failure in computer networks. For example, with regard to idea spreading in a social network, the threshold function ϕ could be more like a random function, unless the idea is surprising or causing people to be angry. For the biological virus spreading in human society, usually, most people have the same or similar resistance to a virus. In this case, the uniform threshold function better captures the cascading model of virus spreading. Many real-world cascading procedures include either the random threshold function or the uniform threshold function. However, there are cascading failures that are neither the random nor the uniform threshold function. Considering a *super virus* for example, it infects all its neighbors immediately in one step, just as the forest fire burns in a forest. In this case, the cascading procedure is not captured by either the random threshold or the uniform threshold. However, we remark that the random threshold function and uniform threshold function suffice for us to investigate the principles of security of networks, and that these principles are essential or even sufficient for us to solve the security of networks with an arbitrarily given threshold function.

To understand the relationship between the two strategies of physical attacks and cascading failure models of attacks, we introduce the notion of *injury set* of physical attacks.

Definition 2.4. **(Injury set.)** Let $G = (V, E)$ be a network and S be a subset of V. The physical attacks on S are to delete all nodes in S from G. We say that a node v is injured by the physical attacks on S if v is not connected to the largest connected component of the graph obtained from G by deleting all nodes in S.

We use $\text{inj}_G(S)$ to denote the injury set of S in G.

In [29], it is shown that the infection sets are much larger than the injury sets of attacks of the same sizes in the networks of the existing models, including our security model. In fact, this phenomenon is the result of the thesis of [44]: that the second best strategy is the attack through diplomacy–the attack by cascading failure of virus, and that the third best strategy is the attack of the enemy's army–the physical attack of deleting nodes or edges.

To understand this result, we introduce two experiments on the networks of the ER model and the PA model.

The first model is the ER model [15, 16]. In this model, we construct a graph as follows: Given n nodes $1, 2, \ldots, n$, and a number p for any pair i, j of nodes i and j, we create an edge (i, j) with probability p.

We depict the curves of sizes of the infection set and the injury set of attacks of top-degree nodes of networks of the ER model in Figures 1(a) and b.

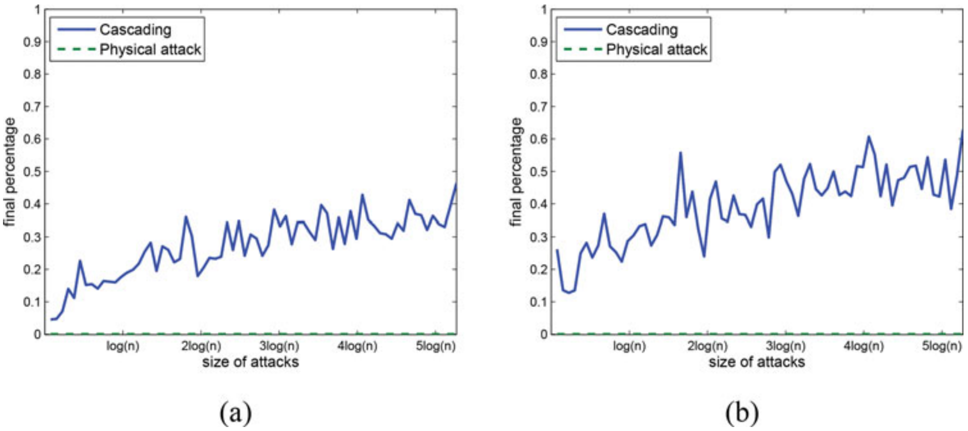

Figure 1 Graphs (a) and (b) are the curves of fractions of sizes of infection sets and injury sets by attacks of the top-degree nodes of small sizes, i.e., up to $5 \log n$, for networks of the ER model for $n = 10,000$ and for $d = 10$ and 15, respectively. The sizes of the infection sets are the largest ones among 100 times of attacks under the random threshold cascading failure model. The infection sets and injury sets correspond to the solid and dashed curves, respectively.

The second is the PA model [4]. In this model, we construct a network by steps as follows: At step 0, choose an initial graph G_0. At step $t > 0$, we create a new node, v say, and create d edges from v to nodes in G_{t-1}, chosen with probability proportional to the degrees in G_{t-1}, where G_{t-1} is the graph constructed at the end of step $t - 1$, and d is a natural number.

We depict the comparisons of sizes of infection sets and injury sets of attacks of the top-degree nodes of networks generated from the PA model in Figures 2(a) and 2(b).

Figures 1(a), 1(b), 2(a), and 2(b) show that for any network, G say, generated from either the ER model or the PA model, the following properties hold:

(1) The infection sets are much larger than the corresponding injury sets.

This means that to build our theory, we need to consider only the attacks of cascading failure models.

(2) The attacks of top-degree nodes of size as small as $O(\log n)$ could cause a constant fraction of nodes of the network to be infected under the cascading failure models of attacks.

This means that the networks of the ER and PA models are insecure for attacks of sizes as small as $O(\log n)$.

Remarks: (i) We notice that this discovery is an observation of our experiments with particular choice of parameters in Figures 1 and 2. The discovery is sufficient for us to propose the correct definition of security of networks. Clearly, experimental proofs are sufficient for the proof of insecurity of networks. From this result, we know that the networks of the ER and the PA models are never secure. (ii) Our experiments in Figures 1 and 2 actually predict some general results of quantitative insecurity of the networks of the models. However, we leave it as an open question here.

From Figures 1(a), 1(b), 2(a), and 2(b), we have the following results:

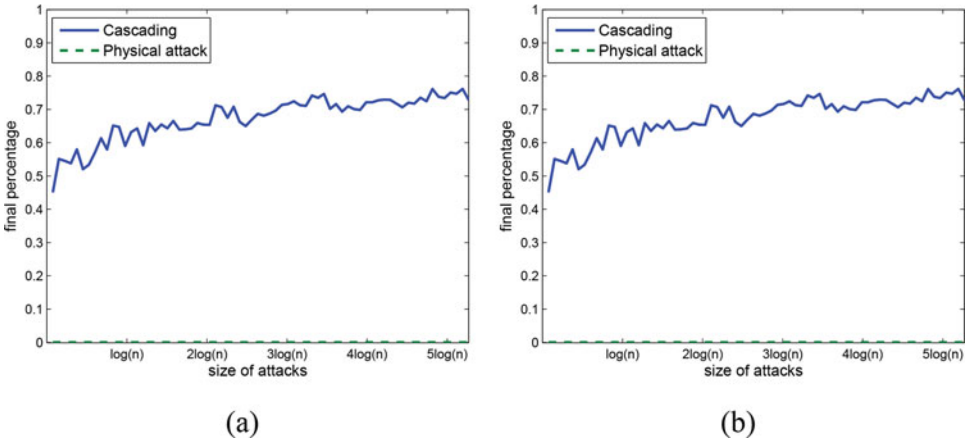

Figure 2 Graphs (a) and (b) are the curves of fractions of sizes of infection sets and injury sets by attacks of the top-degree nodes of small sizes, i.e., up to $5 \log n$, for networks of the PA model for $n = 10,000$ and for $d = 10$ and 15, respectively. The sizes of the infection sets are the largest ones among 100 times of attacks under the random threshold cascading failure model. The infection sets and injury sets correspond to the solid and dashed curves, respectively.

(1) The main issue of network security is to resist the global failure of networks under cascading failure models;
(2) For both theory and applications, it suffices to guarantee the security against attacks of sizes polynomial in $\log n$; and
(3) Topological structures of networks are essential to the security of the networks, observed from the comparison of infection fractions between the ER and the PA models.

(1) and (3) are obvious. For (2), why is poly $\log n$ size of attacks sufficient for a security theory? The reasons are: (i) the networks of the classical models are insecure for attacks of $O(\log n)$ sizes, (ii) a security theory of resisting cascading failures of poly $\log n$ attacks requires new ideas different from all the existing models, and (iii) in practice, the number of nodes of a real-world network is bounded by a large constant, N say, in this case, by appropriately choosing the exponent hidden in the poly $\log N$, the security of the network resisting cascading failures of the poly $\log N$ attacks is sufficient for us to guarantee the security of the network in real-world applications.

3. DEFINITIONS OF SECURITY

As mentioned in Section 2, the main issue of network security is the security for cascading failure models and for attacks of sizes polynomial in $\log n$, i.e., of sizes $\log^{O(1)} n$, where $O(1)$ is a constant.[1]

We propose mathematical definitions for security networks based on the analysis in Section 2.

[1] This is different from $\log n$, a reader may not confuse. It allows $O(1)$ to be an arbitrarily chosen constant.

Our goal here is to explore the principles of network security by constructing a family of networks with various sizes such that sufficiently large networks in the family are provably secure with probability arbitrarily close to 1.

We consider the security of a family of networks of all sizes, $\{G_n\}$ say. We define the security and robustness of networks under the threshold cascading failure model as follows. Let $G = G_n$ and n be the number of nodes of the network G. We first give informal definitions of security and robustness.

Informal definition of security: With probability $1 - o(1)$,[2] the following event occurs: for any initial set S of size $\log^{O(1)} n$, S will not cause a global cascading failure, that is, the size of the infection set of S in G is $o(n)$.

Informal definition of robustness. With probability $1 - o(1)$, a small number, i.e., $\log^{O(1)} n$, of random choices of the initial set S will not cause a *global cascading failure*, that is, the size of the infection set of S in G is $o(n)$.

Let $\mathcal{M}$ be a model of networks. We investigate the security of networks constructed from model $\mathcal{M}$. We define the security of networks for attacks of cascading failure with both random threshold and uniform threshold, respectively. Suppose that G is a network of n nodes, constructed from model $\mathcal{M}$, for large n.

Definition 3.1. **(Random threshold security.)** We say that the networks of model $\mathcal{M}$ are secure resisting cascading failure with the random threshold function if, for every constant $c \geq 1$ and each small constant $\epsilon < 1$, there is a natural number N such that for every $n > N$ for every network G of n nodes generated by model $\mathcal{M}$, with probability $1 - o(1)$, the following holds:

For any set S of size bounded by $\log^c n$, the size of the infection set (or cascading failure set) of S in G is less than or equal to $\epsilon \cdot n$.

Definition 3.2. **(Uniform threshold security.)** We say that the networks of model $\mathcal{M}$ are secure resisting cascading failure with uniform threshold if, for an arbitrarily given small number ϕ, i.e., $\phi = o(1)$, constant c, and small constant ϵ, there is a natural number N such that for any $n > N$, and for any network G of size n generated by model $\mathcal{M}$, with probability $1 - o(1)$, for any set S of size bounded by $\log^c n$, the size of the infection set of S in G, written by $\inf_G^{\phi}(S)$, is bounded by $\epsilon \cdot n$.

Definition 3.3. **(Security of model $\mathcal{M}$.)** Let $\mathcal{M}$ be a model of networks. We say that model $\mathcal{M}$ is secure if networks constructed from model $\mathcal{M}$ are secure for both random and uniform threshold cascading failure models of attacks.

The notion of robustness can be established similarly, based on the informational definition of robustness, with which a theory of robustness of networks maybe built.

In Definitions 3.1 and 3.2, the sizes of attacks or random errors are polynomial in $\log n$, the reasons for which are analyzed in Section 2.

In Definitions 3.1 and 3.2, we notice that both security and robustness of a family of graphs depend on a threshold function $\{\phi(v) \mid v \in V\}$.

[2] $o(1)$ is a sequence, a_n say, such that $\lim\limits_{n\to\infty} a_n = 0$ instead of a fixed number; $o(n)$ is a function f such that $\lim\limits_{n\to\infty} \frac{f(n)}{n} = 0$. The notations are standard in mathematics and are assumed throughout the article.

Our notions in Definitions 3.1 and 3.2 consider the most natural distributions $\{\phi\}_{v\in V}$, i.e., the random distribution and the uniform distribution. By using these definitions, we are able to fully explore the roles of network structures in the security of networks.

A reader may wonder: Can we define the notions of “absolute" security of a graph such that it is independent of any threshold function $\{\phi(v)\}_{v\in V}$? To answer this question, let us consider a special case. For a graph $G = (V, E)$, for every $v \in V$, let $\phi(v) = \frac{1}{d_v}$ where d_v is the degree of v in G. In this case, we have:

(1) If G is connected, then an attack on any single node v will cause the infection of the whole graph.

In this case, any infected node carries a super virus, which immediately infects the whole network through all the links. Therefore, with the given threshold function, the security of G is impossible to achieve.

(2) By (1), the only way to guarantee the security of G is to remove some nodes or edges of G or impose some “controllers" on certain nodes so that the network consists of small connected components, for which different connected components are isolated among each other.

This observation implies that, in this case, the security strategy of G with the given threshold function actually corresponds to a strategy of physical attacks or “physical controlling.”

The results in (1) and (2) establish a relationship between physical attack and super virus spreading. This means that physical attack corresponds to the controlling of super virus spreading. However, in the present study, we focus on the security theory with the random threshold and uniform threshold functions.

We remark that the security of networks by physical attacks and super virus spreading is different from our theory and independently interesting. Although, we believe, our theory has provided a foundation for solving the problems of security by physical attacks and super virus attack.

In addition, we remark that it is interesting to study the security of networks with other interesting threshold functions, such as for each $v \in V$, $\phi(v) = \frac{k}{d_v}$ for some $k > 1$, where d_v is the degree of v in G. For this question, combinatorial properties of G might play more important roles in the security of G. Here, we leave it as an open question.

4. EXISTING MODELS FAIL TO GENERATE SECURE NETWORKS

A reader may wonder, if there have already been a number of models of networks, in which some of them may generate secure networks. We will verify that this is not the case. For this, we briefly introduce the existing models. We have already studied the ER and the PA models, so we study the other models here.

(1) The copying model:

Both linear growth copying and exponential growth copying models have been proposed [22]. For simplicity, we will use only the linear growth copying model with a slight modification[3] for the first edge of a newly created node. The model has two parameters: a *copying factor* $\alpha \in (0, 1,)$, and a number d, which is the average number of edges of the graphs generated by the model. At each step, one node v is created and v is then given d out-links. Suppose that G_t is constructed. At step $t + 1$, we create a new node and randomly and uniformly choose a node u in G_t, with which we create a directed edge from v to u, and u is called the first out-going neighbor of v. For each $i \in \{2, \ldots, d\}$, with probability α, we randomly and uniformly choose a node x_i in G_t, in which case, create a directed edge from u to x_i, and x_i is called the ith out-going neighbor, otherwise, let y_i be the ith outgoing neighbor of u, in which case, call y_i the ith outgoing neighbor of v, and create a directed edge from v to y_i.

The directions are used to guide the construction of graphs. After the construction, we regard the graph as undirected. By definition, a newly created node, v say, first links to a randomly chosen node u, and then, with certain probability, links to the first group of neighbors of u (i.e., the neighbors of u created at the same time when u was created). This is the essential idea of the copying model, which makes sure that the generated graphs are rich in interesting subgraphs such as connected union of k-cliques for $k \geq 3$.

The most remarkable fact is that the copying model generates power-law graphs, the same as the PA model. The reason is that PA does not appear explicitly in the copying model. However, it does appear implicitly by randomness and local rules, because a newly created node will have more links to the first group of neighbors of its first random neighbor. The model demonstrates that randomness and local rules generate power-law graphs, and that graphs generated by the copying model are rich in interesting subgraphs, interpreted as communities.

Therefore, the copying model introduced the idea of copying, which is different from the PA, and which has been used in many other models, including those we will introduce following.

(2) Random walk model:

The random walk model [46] simulates the random walk behavior of friends discovery in online social networks. Each new node performs a random walk starting with a randomly chosen node and links to certain nodes appearing in the random path. We will use a modified version of the model.

Our modified version proceeds by steps as follows. Suppose that G_t has been defined. At step $t + 1$, a new node, v say, is created, then randomly pick a node u_1 from G_t and create an edge between v and u_1. For each $i = 2, \ldots, d$. Suppose that u_j is defined, then perform one step from u_j to u_{j+1}, with probability $p_{u_{j+1}}$, create an edge between v and u_{j+1}, in which case, increase i by 1, otherwise, increase j by 1. In our modification, we choose $p_u = \frac{1}{d_u}$ for all u's, where d_u is the degree of u. Therefore, the step from u_j to u_{j+1} is a step of random walk with uniform distribution.

(3) Nearest neighbor model:

This model uses the clustering effect as rules in the creation of edges [46]. We will use a modified version of the model. It proceeds as follows. Suppose that G_t is defined. At step $t + 1$, we create a new node, v say, then randomly and uniformly choose a node u in G_t and create an edge between v and u. In addition, we randomly create $d - 1$ edges between the neighbors of u's.

[3] The modification here and elsewhere for the other models is to make sure that the networks generated by the different models can be compared. Of course, in each of the modifications, we kept the essential idea of the different models so that the experimental results of the networks of the models essentially represent the results of the networks of the models.

(4) Random surfer model:

The random surfer model was proposed in [5]. It proceeds as follows. Given d and a probability p, at step $t + 1$, create a new node, v say, and repeat the following steps d times:

Step 1. Let u be a randomly and uniformly chosen node in G_t. Set $x = u$.
Step 2. With probability p, create an edge between v and x, and stop.
Step 3. Otherwise, perform a one-step random walk from x to a node y, set $x \leftarrow y$ and go back to Step 2).

(5) Duplication model: The model has a selection probability p. We use the version described in [10]. It proceeds as follows. Suppose that G_t is defined. At step $t + 1$, create a new node, v say, randomly and uniformly choose a node u in G_t, create an edge between v and u, and for every neighbor w of u, with probability p, create an edge between v and w.

We omitted a number of models. The model in [15] generates random graphs, which was studied in Figures 1(a), 1(b). The small-world model [48] generates small-world graphs, which are clearly insecure by simply attacking on a small number of consecutive nodes. We omitted the following models:

(1) The forest fire model [25]:

This model uses the idea: copying edges to generate dense graphs. Because we have the copying model, and we are concerned mainly with sparse graphs (anyway, dense graphs are much less frequent in the real world).

(2) The model for social graphs in [45]:

This model was proposed to better capture the communities and high clustering coefficients in social networks. The model is basically an extension of the copying model such that a new node links to a randomly picked node and then links to a certain number of nodes within a distance of two of the randomly chosen node. The graphs generated from this model do have more nodes in a connected union of cliques and triangles. We omitted it because the same ideas already appeared in the models we selected.

(3) The Kronecker model [24]:

The Kronecker model is a recursive generator by using a tensor product, which is specific and for which the difficulty is to choose the initial module for the tensor products.

(4) The so-called dK-graphs [33]:

These so-called graphs simulate given graphs better and better with increasing d, which is not suitable for us. Finally, we notice that it is only reasonable to compare the networks with the same sizes and the same number of edges to explore exactly the differences of structures. This is also the reason why we omitted some of the reasonable models such as the forest fire model and the Kronecker model.

In Figure 3, we depict the curves of sizes of infection sets of attacks of $\log n$ top-degree nodes on networks generated by the copying model, the random walk model, the nearest neighbor model, the random surfer model, and the duplication model. All the networks have average number of edges $d = 5$. The parameter α in the copying model is 0.3. The networks have sizes ranging from $1,000$ to $10,000$. Each of the curves corresponds to the largest infection sets among 100 times of attacks, for which each time, the thresholds $\phi(v)$'s

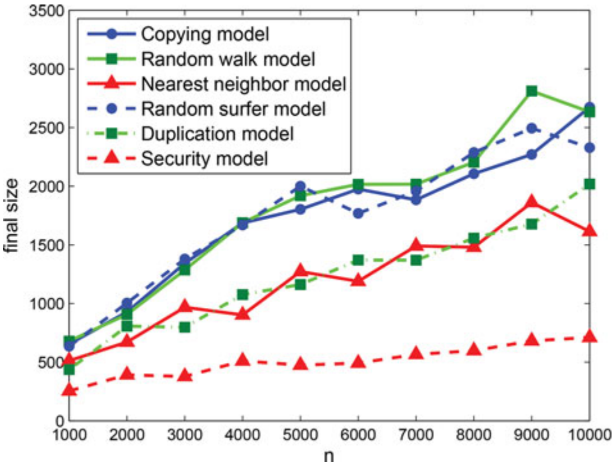

Figure 3 Security curves of networks of the copying model, the random walk model, the nearest neighbor model, the random surfer model, the duplication model and our security model. In this figure, we consider the case of random thresholds. It depicts the curves of the greatest size of the cascading failure sets of attacks of 100 times of size $\log n$ for each n less than or equal to 10, 000. The curves describe the greatest sizes of the final cascading failure sets among 100 times of attacks of the random thresholds. The sizes of the initial attacks are always $\log n$, where n is the number of nodes of the networks.

are randomly and uniformly distributed. By observing the curves of the existing models except for our security model, in Figure 3, we have that for the random threshold cascading failure model, networks of the nearest neighbor model have the best security performance, the networks generated from the duplication model have the worst security performance, and for each of the copying model, the random walk model, the nearest neighbor model, the random surfer model, and the duplication model, it is possible to have a cascading failure of constant fraction of nodes by attacking $\log n$ many nodes top-degree nodes.

From Figures 1(a), 1(b), 2(a), and 2(b), we know that nontrivial networks of the existing models, including the ER model, the PA model, the copying model, the random walk model, the nearest neighbor model, the random surfer model, and the duplication model, are insecure.[4] This poses fundamental questions such as: Are there networks with power-law and -world properties that are secure by Definitions 3.1 and 3.2? What mechanisms guarantee the security of networks? What are the principles of security of networks? In the present work, we will answer these questions.

5. SECURITY MODEL OF NETWORKS

To establish the theory of network security, we first examine some high-level philosophical questions: Can large-scale networks be secure? If the answer is yes, where can we find the resources for the security of networks? To understand these questions, we observe

[4] For insecurity, experimental results are sufficient. We notice that the simple strategy of top-degree attacks have already caused a significantly large number of cascading failures of the networks. For better strategies of attacks, the cascading failure sets of the networks of the existing models may be even larger. Therefore, the best possibility we can expect for the networks of the existing models could only be the robustness of the networks, which we leave here as an open question.

that real-world networks are naturally evolving interacting systems in which human behaviors are probably involved and that large-scale engineering networks are generated by both human behaviors and human intelligence. More importantly, our theory is to provide a foundation for guaranteed security of large-scale engineering networks. The arguments indicate that the security of networks could be achieved by both the laws of nature evolving and combinatorics, corresponding to human behaviors and human intelligence, respectively. This can be intuitively understood as "nature secures its networks" and "human intelligence secures networks."[5]

According to these arguments, [29] proposed a security model of networks. It was shown that the networks of the security model are much more secure than the networks constructed from the ER and PA models. The security model reflects the best strategy of the defensive stratagem of the thesis [44] (published in 512 BC); that is, the attack by frustrating the enemy's strategy–the attack of rules of networks, as pointed out in [29].

5.1. The Security Model

The starting point of our theory is the security model in [29], for which we introduce definitions as follows.

Definition 5.1. **(Security model, [29])** Given a homophyly exponent (also called as affinity exponent[6]) $a \geq 0$ and natural number d, we construct a network by stages.

1. Let G_2 be the graph with two nodes and d multiedges such that each node has a distant color and is called a *seed.*

   (Notice that the initial graph G_2 could be an arbitrarily given graph such that each node has a distinct color and is called a seed. This will not change the principle of the networks of the model.)
2. Let $i > 2$. Suppose that G_{i-1} has been defined. Define $p_i = \log^{-a} i$.
3. Create a new node v.[7] With probability p_i, v chooses a new color,[8] c say. In this case,
   (a) we say that v is the seed node of color c,
   (b) (PA scheme) create one edge (v, u), such that u is chosen with probability proportional to the degrees of nodes in the whole graph G_{i-1},[9] and (Notice that we are allowed to create ONLY one edge (v, u) by the PA scheme, here. This property is essential to the security proofs of the networks. In fact, even if the creation of two such edges might cause the networks of the model to be insecure. This limitation of our model, here, is harmless, because in real-world computer networks, there are no such even edges.

[5] It does not make sense, if a network security theory were built independently of both human behaviors and military ideas.

[6] This is our key new idea. It allows us to quantitatively measure the power of homophyly/kinship in naturally evolving networks.

[7] In Darwin's theory [12], it is assumed that the resource is limited, allowing only constantly many individuals' survival to the adulthood. So we create only one node in a single step.

[8] In Darwin's theory [12], any two individuals are distinct, so each individual is special, playing different roles. This step reflects the fact that any two individuals are distinct, one of the facts of Darwin's natural selection.

[9] If all the newly created d edges linking from v to nodes in G_{i-1} are chosen with probability proportional to their degrees, then the model is the homophyly/kinship model [28], which fully explores the principles of natural selection of homophyly/kinship in nature evolving.

However, in social or biological networks, there do exist such edges. Allowing more such edges creates a new type of reciprocity among different communities, which are useful in other applications of networks such as in guaranteeing the emergence of cooperation in evolutionary games in the networks generated by the modified model. Therefore, we may allow more such edges in simulating naturally evolving networks. The restriction of at most one such edge, here, is required only by the security of the computer network. For other engineering networks such as networking computing systems, or social media networks, such edges might even be very useful.)

(c) (Randomness) create $d-1$ edges (v, u_j), $j = 1, 2, \ldots, d-1$, where u_j's are chosen randomly and uniformly among all seed nodes in G_{i-1}.[10]

4. (Homophyly[11] and preferential attachment) otherwise. Then v chooses an old color, in which case,

(a) let c be a color chosen randomly and uniformly among all colors in G_{i-1},

(b) define the color of v to be c, and

(c) create d edges (v, u_j), for $j = 1, 2, \ldots, d$, where u_j's are chosen with probability proportional to the degrees of nodes in the whole graph G_{i-1} among all the nodes that have the same color as v in G_{i-1}.

It is clear that Definition 5.1 is a dynamic model of networks for which homophyly, randomness, and PA are the underlying mechanisms.

As shown in [29], networks constructed from the security model are much more secure than those of the ER and PA models. To understand the intuition of the security model, we use a figure from [29], Figure 4, here. It depicts three curves of the sizes of the infection sets of attacks of top-degree nodes of sizes up to $5 \cdot \log n$ under the random threshold cascading failure model on networks generated from the security model, the ER model, and the PA model, respectively. The curves correspond to the largest infection set among 100 times of attacks over random choices of thresholds of the networks. The figure shows that networks of the security model are indeed much more secure than those of both the ER and the PA models, even if we just take the homophyly exponent (or affinity exponent) $a > 1$ in the security model.[12]

Figure 3 shows that the networks generated by our security model are more secure than the networks generated by all the existing models.

[10] This step ensures that the induced subgraph of all the seed nodes of a network of the model is an expander, allowing easy and quick interacting among the nodes of the whole network. Therefore, this step determines a very nice global property of the networks of the model.

[11] In Darwin's theory [12], kinship is the means of individual fitness and the intrinsic mechanism of social groups in the evolution of species. Our notion of homophyly is an extension of the idea of Darwin's kinship. It simulates the formation of social groups in Darwin's evolution of species, and it determines the local structures or natural communities of the networks of the model.

[12] This experiment shows that the security model generates networks of practical scales with better performance for the security of networks. It also indicates that there is much room for research on the effect and roles of changing the affinity exponent a for the networks of the security model. Research on the roles of changing a could explore new principles such as the principle of self-organization of individuals in nature and society, the principle of natural community structure of networks, the practical principle and criteria for security and robustness of networks, the principle for the emergence of cooperation in evolutionary games in networks, and the principle of social morphogenesis.

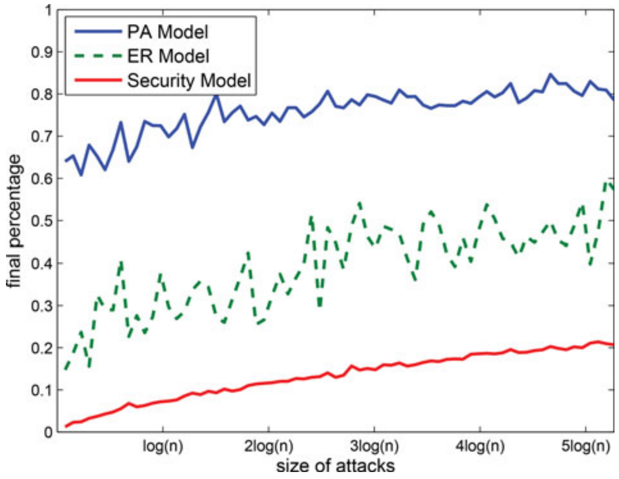

Figure 4 The curves are cascading failures of networks of the ER model, the PA model, and the security model for $n = 10,000$, $d = 15$ and $a = 1.5$.

5.2. Basic Observations of the Networks of the Security Model

The experiments in [29] verified the following properties of the networks generated by our security model:

(1) The mechanisms of homophyly, randomness, and PA ensure that the networks of the security model satisfy a number of structural properties such as

(a) (Small-community phenomenon) A network, G say, is rich in quality communities of small sizes.

In fact, let S be a maximal homochromatic set of G. Then, the induced subgraph of S, written by G_S, is highly connected, and the conductance of S, written by $\Phi(S)$, is bounded by a number inversely proportional to a constant power of the size of the community, i.e., less than or equal to, $O(\frac{1}{|S|^{\beta}})$, for some constant β, where $|S|$ is the size of S.

(b) (Internal centrality) Each community is the induced subgraph of nodes of the same color, which follows a preferential attachment, and hence has only a few nodes dominating the internal links of the community.

This shows a remarkable local heterogeneity of the networks.

We call the induced subgraph of a maximal homochromatic set of G a *natural community* of G.

(c) (External decentrality) For a natural community X, the set of external neighbors of nodes of X are homogenously distributed among different natural communities.

(2) (Power law) The networks follow a power law.

(3) (Small-world property) The networks have small diameters.

(4) (Robustness) Most nodes in a natural community have neighbors only within their own community.

(5) (Stability) For a node v, either seed or nonseed, the degree of v is contributed largely by nodes of v's own community.

(6) (Inclusion/exclusion property) There are no edges between the nonseed nodes of distinct natural communities.

(7) A nonseed node x in a community G_X, created at time step t can be infected by nodes in a distinct community G_Y, only if the seed node y_0 of G_Y is created at a time step $s > t$ and an edge (y_0, x) is created at step s by 3(b) of Definition 5.1.

The structural properties in (1) allow us to develop a methodology of community analysis of networks; (2) and (3) show that the networks constructed from the security model have the most important properties of usual networks; (4) and (5) ensure that nodes in a natural community are difficult to infect by the external neighbors of the community, so that almost all the natural communities are strong; and (6) and (7) ensure that infections among different communities are hard. These observations intuitively explain the reason why networks constructed from the security model show much better security than that of the classic ER and PA models.

The arguments in the list imply that the small-community phenomenon, local heterogeneity, global homogeneity, and global randomness are essential to the security of networks with power law and small-world property.

In this study, we will show that the security model is provably secure by Definition 3.3. The key idea of the proofs is a merging of some principles of the natural selection of homophyly and combinatorics.

Remarks: (i) As mentioned, the security model of networks uses the ideas from both Darwin's natural selection and combinatorics. (ii) Engineering networking involves the laws of both nature evolving and human intelligence. (iii) Security is only one of the many important requirements of real networks, although it is extremely important. (iv) In the real world, an engineering network must be built such that it is secure, and such that people are comfortable, easy, and quick to use the network. (v) As a new engine, an engineering network must satisfy many engineering requirements together with the security requirement to prompt production. This poses new issues for engineering networks.

5.3. The Security Model Truthfully Reflects Naturally Evolving Networks

The security model assumes that homophyly/kinship is the mechanism of social groups of the networks. This truthfully reflects Darwin's idea that kinship is the individual fitness for cooperation in social groups, which is essential to the survival of the species in the evolution of species. This observation indicates that the security model actually simulates many of the networking systems in nature and society. It was verified [26, 27] that homophyly is indeed the mechanism of natural communities in real-world networks, for which the security model provides the theoretical resources for analyzing the natural communities of real-world networks. Furthermore, [32] proposed the notion of structure entropy of networks to quantitatively measure the nondeterminism of interactions and communications in networks, and proposed greedy algorithms to find the partitions and partitioning trees of networks by minimizing the two- and three-dimensional structure entropy of networks. By using these algorithms with the cell sample networks of five cancers constructed from the gene expression profiles, we found the tumor modules and submodules such that (i) a module or submodule found by our algorithms is a type or subtype defined in cancer biology, and (ii) each module or submodule found by our algorithms is defined by a unique gene expression pattern, giving rise to a three-dimensional, high-definition gene map of cancer types, for the first time. More importantly, clinical data analyses demonstrate that the patients of a submodule found by our algorithms share similar survival times, survival indicators, and International Prognostic Index (IPI) scores, and that patients of distinct submodules found by our algorithms are remarkably different in overall survival times, survival ratios, and IPI scores. Our definition of tumor subtypes is purely the output of our algorithms. Our classifications of tumor types are the first such results with interpretable and distinguishable clinical performances. We remark that such kind of cancer classifications had never been achieved in existing biotechnology.

From such progress, we know that homophyly is the semantical principle of natural communities in networks, which our security model has fully explored, and that structure entropy minimization is the syntactic principle of natural communities and natural hierarchical structures of networks, for which a theory of structure entropy of graphs may be established. Our results also indicate that the minimization of nondeterminism of structures of a network is the principle of self-organization of the individuals of the network, which would be a fundamental principle for network analysis.

Therefore, our security model not only generates secure networks, but also truthfully reflects some of the fundamental principles of naturally evolving networks.

6. THE MAIN THEOREMS AND OVERALL IDEAS

We use $\mathcal{S}(n, a, d)$ to denote the set of random graphs of n nodes constructed by the security model with homophyly exponent a and average number of edges d.[13]

Let G be a network constructed from the security model. We have that each node is assigned a color. This new dimension of colors allows us to characterize the structures of the networks. In our security model, every node has its own characteristics from the very beginning of its birth. This feature is remarkably different from the classic models such as the ER and the PA models. Anyway, the extra dimension of colors is essential to our understanding of security of networks.

We call a maximal set of nodes of the same color, κ say, a *homochromatic set*, written by S_κ. We call a homochromatic set or the induced subgraph of a homochromatic set of G a *natural community of* G.

We say that an edge is a *local edge* if two of its endpoints share the same color, and *global edge*, otherwise.

First, we prove some structural properties of networks of the security model.

Theorem 6.1. (Fundamental theorem of the security model.) *Let $a > 1$ be the homophyly exponent and $d \geq 4$ be a natural number. Let $G = (V, E)$ be a network constructed by $\mathcal{S}(n, a, d)$.*

Then, with probability $1 - o(1)$, the following properties hold:

(1) (Basic properties):

(i) (The number of seed nodes is large) The number of seed nodes is bounded in the interval $[\frac{n}{2\log^a n}, \frac{2n}{\log^a n}]$.

[13] In both the PA model and the security model in Definition 5.1, we consider d as a constant. Thus, in all notations of $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, and $\omega(\cdot)$ throughout the article, d is always absorbed.

(ii) (Natural communities are small) Each homochromatic set has a size bounded by $O(\log^{a+1} n)$.

(2) (Self-organizing law or holographic law) For degree distributions, we have

(i) (Internal centrality) The degrees of the induced subgraph of a homochromatic set follow a power law.

(ii) The degrees of nodes of a homochromatic set follow a power law.

(iii) (Power law) Degrees of nodes in V follow a power law, that is, there is a constant β*, such that the number of degree k nodes in the network is proportional to* $\frac{1}{k^\beta}$.

(iv) The power exponent of the power laws in (i)–(iii) are all the same.

(3) For node-to-node distances, we have

(i) (Local-communication law) The induced subgraph of a homochromatic set has a diameter bounded by $O(\log\log n)$.

(ii) (Small-world phenomenon) The average node-to-node distance of G is bounded by $O(\log n)$.

(iii) (Algorithmic small-world phenomenon) There is an algorithm to find a short path between two arbitrarily given nodes in time $O(\log n)$.

(4) (Small-community phenomenon [30, 31]) There are $1 - o(1)$ *fraction of nodes of G, each of which belongs to a homochromatic set, W say, such that the size of W is bounded by* $O(\log^{a+1} n)$*, and the conductance of W,* $\Phi(W)$*, is bounded by* $O(\frac{1}{|W|^\beta})$ *for* $\beta = \frac{a-1}{4(a+1)}$.

This shows that the network is rich in quality communities of small sizes.

Theorem 6.1 explores an interesting topology of a network G: (i) G consists of a local structure and a global structure, (ii) the local structure of G is determined by the small communities that have a number of local properties, and (iii) the global structure of G follows its own laws. The network is rich in quality communities of small sizes, which compose the interpretable local structures of the network. To the contrary, there is a global structure of the network that ensures that the whole network is highly connected, with a power-law distribution and a small-diameter property. Communications in G have two types: (i) local communications within the small communities of length $O(\log\log n)$, and (ii) global ones, which make the whole network to be highly connected in length $O(\log n)$. More importantly, there exists a *local algorithm* running in time $O(\log n)$ to navigate in the whole network. Most of the communications are local ones having length within $O(\log\log n)$, and the rest of the communications are global ones with length bounded by $O(\log n)$. The construction of a network with explicit marks of local and global structures by Definition 5.1 allows local algorithms of time complexity $O(\log n)$ to find useful information in the whole network. This suggests a new algorithmic problem, that is, to find network algorithms of time complexity polynomial in $\log n$ for finding useful information.

Theorem 6.1 ensures that all the communities are small. This guarantees that even if a single node in a small community infects the whole community, the cascading failure is still a local cost. However, it is not intuitive to understand from Theorem 6.1 the reason why networks of the security model are secure. In fact, to prove the security theorems, we need to develop some combinatorial properties of the networks. In [29], the authors analyzed experimentally some of these properties.

Suppose that $G = (V, E)$ is a network constructed from the security model. For a subset $X \subset V$, we always use G_X to denote the induced subgraph of X in G.

For a set of nodes S, we define $C(S)$ to be the set of colors that appear in S. For a node v, we use $N(v)$ to denote the set of neighbors of v. Given a node v, we define the *length of degrees of* v to be the number of colors associated with the neighbors of v, i.e., $|C(N(v))|$, written by $l(v)$.

Suppose that $N_1, N_2, \ldots, N_l$ are the sets of all the neighbors of v such that nodes in each N_i share the same color, and that nodes in different N_i's have different colors. Let d_i be the size of N_i, for each $i \in \{1, 2, \ldots, l\}$. Suppose that $d_1 \geq d_2 \geq \cdots \geq d_{l(v)}$ (ties break arbitrarily). In this case, we say that d_i is the *ith degree of* v, and the color of nodes in N_i is the *ith color of neighbors of* v, for all $i \in \{1, 2, \ldots, l\}$.

The length of degrees, the ith degree and the ith color of neighbors of vertices have some interesting properties, including those validated by experiments in [29]: (i) the length of degrees of a vertex is always bounded by $O(\log n)$, (ii) the first degrees d_1's are large, (iii) the second degrees are always as small as constants, and (iv) for a vertex v, if the length of degrees of v is $l(v) > 1$, then for any $i > 1$, the ith color of neighbors of v is distributed homogenously. These properties are essential to the experimental analysis of security of the networks in [29].

To theoretically prove the results, we define some useful notations.

Definition 6.2. Let $G = (V, E)$ be a network constructed from the security model. Given a node $v \in V$:

1. For every j, we define the jth degree of v at the end of time step t to be the number of the jth largest set of homochromatic neighbors of v at the end of time step t, written as $d_j(v)[t]$.
2. We define the jth degree of v to be the jth degree of v at the end of the construction of network G, written as $d_j(v)$.
3. We define the length of degrees of v at the end of time step t to be the number of colors associated with neighbors of v at the end of time step t, written as $l(v)[t]$.
4. We define the length of degrees of v to be the length of degrees of v at the end of the construction of G, written as $l(v)$.

In sharp contrast to classic graph theory, for a network constructed from our security model, G say, and a vertex v of G, v has a *degree priority*. This new feature is certainly universal in real networks in the following sense: a natural community is an interpretable group of nodes of a network such that nodes of the same community share common attributes. In this case, a vertex v may have its own community and may link to some neighbor communities by some priority ordering. In our model, a node v more likely contacts nodes with which it shares the same color (or feature), and has not much contact to nodes outside its own community.

Definition 6.3. (Degree Priority.) Let v be a node of G constructed from the security model created at time step t_0, and $t \geq t_0$.

1. Suppose that $N_1, N_2, \ldots, N_l$ are all the homochromatic neighbors of v at the end of time step t listed decreasingly by the sizes of the sets N_j. For $d_j = |N_j|$ for each j, we say that $(d_1, d_2, \ldots, d_l)$ is the degree priority of v at the end of time step t, written as $\vec{d}_v[t] = (d_1, d_2, \ldots, d_l)$.

2. We define the degree priority of v in G to be the degree priority of v at the end of the construction of G, written as $\vec{d}_v$.

The degree priority of nodes in G satisfies some nice probabilistic and combinatorial properties.

Theorem 6.4. (Degree Priority Theorem.) *Let G be a network constructed from the security model with $d \geq 2$, and $a > 1$. Then, with probability $1 - o(1)$, for a randomly chosen node v, the following properties hold:*

(1) The length of degrees of v is bounded by $O(\log n)$, which is an upper bound independent of a.
(2) The first degree of v is the number of neighbors that share the same color as v. This means that the first degree of v is contributed by v's own community.
(3) The second degree of v is bounded by $O(1)$, so that for any possible $j > 1$, the jth degree of v is $O(1)$.
(4) If v is a seed node, then the first degree of v is lower bounded by $\Omega(\log^{\frac{a+1}{4}} n)$.

By (2), (3), and (4) of Theorem 6.4, we understand that for a community G_X induced by a homochromatic set X, the seed node, x_0 say, of X has a large first degree and constant second degree, so that it is unlikely to be infected by all the nodes in a single neighbor community, G_Y say. Combining with (1), this ensures that for properly chosen a, the seed node x_0 of G_X is resistant to infection by the collection of all its neighbors outside its own community. Such a community is regarded as a *strong community*. Theorem 6.1 ensures that, for a properly chosen a, almost all of the communities are strong, so that each of them is resistant to infection by the collection of all its neighbors outside its own community.

Combining Theorem 6.1 and Theorem 6.4 gives us a better understanding of the reasons why networks of the security model are secure. However, to prove the security theorems, we have to understand the cascading behaviors of attacks in the networks.

We say that a community, G_X say, is created at time step t if the seed node x_0 of X is created at time step t.

To understand the cascading behaviors, we provide the following definition

Definition 6.5. Let x and y be two nodes of G, and Y be a set of nodes such as $x \notin Y$.

(i) We say that x injures y, if the infection of x contributes to the probability that y becomes infected, or equivalently, there is an edge between x and y. Otherwise, we say that x fails to injure y.
(ii) We define the injury of x from y to be the number of edges between x and y.
(iii) We define the injury of x from Y to be the number of edges between x and the nodes in Y.

We will show that the infection of a community from a neighbor community satisfies a number of combinatorial properties.

Theorem 6.6. **(Infection-Inclusion Theorem.)** *Suppose that X and Y are two maximal homochromatic sets, and that G_X and G_Y are two communities. Let x_0, and y_0 be the seed nodes of X and Y, respectively. Suppose that x_0 and y_0 are created at time step s and t, respectively. Then the injury of G_Y from community G_X satisfies the following properties:*

(1) If $s < t$, then
  (i) The community G_X fails to injure any nonseed node in the community G_Y.
  (ii) The injury of the seed node y_0 of G_Y from the whole community G_X is bounded by a constant $O(1)$.

(2) If $s > t$, then
  (i) All the nonseed nodes in G_X fail to injure any node in the community G_Y.
  (ii) The injury of the seed node y_0 of G_Y from the community G_X is bounded by 1.
  (iii) The injury of a nonseed node in the community G_Y from the seed node x_0 of G_X follows the edge created by Step 3(b) of Definition 5.1.

(3) The seed node y_0 of G_Y can be injured only by
  (i) Communities created at time step $< t$.
  (ii) The seed nodes of communities created at time step $> t$.

(4) A nonseed node y of G_Y can be injured only by seed nodes created at time step $> t$ through the edge created by 3(b) of Definition 5.1.

(1), (2), and (3) of Theorem 6.6, together with Theorem 6.4, show, furthermore, that a seed node, v say, of G is strong against infections from the collection of all the communities other than its own community.

Suppose that X, Y, and Z are three maximal homochromatic sets created at time steps t_1, t_2, and t_3, respectively. Let x_0, y_0, and z_0 be the seed nodes of X, Y, and Z, respectively. It is possible that x_0 infects a nonseed node y_1 of Y, y_1 infects all nodes in Y, including y_0, and y_0 infects a nonseed node z_1 of Z. Theorem 6.6(4) ensures that $t_1 > t_2 > t_3$, and that the edges (x_0, y_1) and (y_0, z_1) must be created by 3(b) of Definition 5.1. The key point is that the edges (x_0, y_1) and (y_0, z_1) must be embedded in a tree of height $O(\log n)$, which we will call the *infection priority tree T* of G.

Precisely, we have:

Definition 6.7. **(Infection priority tree T.)** Let G be a network constructed by Definition 5.1. We define the infection priority tree T to be a directed graph as follows:

1. Let H be the graph obtained from G by deleting all the edges constructed by 3(c) of Definition 5.1, keeping the directions in G.
2. Let T be the directed graph obtained from H by merging each of the maximal homochromatic sets into a single node.

For the infection priority tree T of G, we have the following theorem.

Theorem 6.8. **(Infection Priority Tree Principle.)** *With probability $1 - o(1)$, the following hold:*

1. *The infection priority tree T is a directed tree.*
2. *The height of the infection priority tree T is $O(\log n)$.*

Given a network G generated by the security model, let X be a natural community. Suppose that ϕ is a threshold function of the vertices of G. We say that X is a strong community if $\frac{d'(x_0)}{d(x_0)} < \phi(x_0)$ holds, where x_0 is the seed node of X, $d'(x_0)$ is the number of edges between x_0 and the nodes outside X, $d(x_0)$ is the degree of x_0, and $\phi(x_0)$ is the threshold of x_0.

Therefore, if X is a natural community with seed x_0 and is a strong community, then x_0 cannot be infected by the nodes outside its own community X alone in the sense that if there is no node in X that has been infected, then x_0 cannot be infected by the collection of its neighbors outside X.

By appropriately choosing the affinity exponent a, we are able to show that for a network G of the security model, almost all the natural communities of G are strong communities.

We will see that Theorem 6.6 ensures that the infection of a strong community from a strong community is directed along a path in the infection priority tree, and pointing to the root of the infection priority tree; that Theorem 6.8 ensures that the height of the infection priority tree is almost surely bounded by $O(\log n)$, which is short; and that Theorems 6.1 and 6.4 ensure that almost all the natural communities are strong.

Therefore, Theorems 6.1, 6.4, 6.6, and 6.8 allow us to establish some security theorems of the networks of our security model.

The framework of proofs of our security theorems is as follows. By Theorem 6.8, a targeted or infected strong community triggers at most $O(\log n)$ many strong communities to be infected. By Theorem 6.1, each community has size at most $O(\log^{a+1} n)$. For any initial set of attacks S of size polynomial in $\log n$, suppose that every community that is not strong has already been infected by attacks on S, automatically. Let K be the number of communities that are not strong. Then there are at most $|S| + K$ strong communities that trigger infections in the infection priority tree T. This shows that there are at most $O((|S| + K) \cdot \log n)$ communities, in each of which there is at least one node infected by attacks on S. In this case, again by Theorem 6.1, even if all the nodes in an infected community are infected, the total number of infected nodes is a negligible number compared with the size of the network. This sketch depends on an estimation of K, the number of communities that are not strong, which will be given in the full proofs in later sections.

Therefore (1), (2), and (4) of Theorem 6.6 ensure that the infection of a nonseed node, v say, is always oneway from a seed node created late than v, following an edge in the infection priority tree. By modulo the injury among the seed nodes, we are able to show that the infections of nonseed nodes can proceed only in the infection priority tree of height $O(\log n)$.

Now we fully understand that the combination of Theorems 6.1, 6.4, 6.6, and 6.8 does allow us to prove some security theorems of the security model. This also explores the following security principle of networks.

Security Principle:

1. Small community phenomenon (by Theorem 6.1)
2. The number of seed nodes or hubs is large (by Theorem 6.1)
3. Almost all seed nodes (or hubs) are strong against infections from the collection of all their neighbor communities alone (by Theorem 6.4)

4. There exists an infection priority tree T of G such that infection of nonseed nodes of a community from a neighbor community can only be triggered by seed nodes of the neighbor community through edges in the infection priority tree T of G (by Theorem 6.6)
5. The infection priority tree T of G has height $O(\log n)$ (Theorem 6.8)

Now we are ready to state our security theorems.

By applying Theorems 6.1, 6.4, 6.6, and 6.8, we are able to prove that networks constructed from the security model are secure against any attacks of small sizes under both uniform and random threshold cascading failure models.

For the uniform threshold cascading failure model, we have:

Theorem 6.9. (Uniform threshold security theorem.) *Let G be a graph constructed from $\mathcal{S}(n, a, d)$ with $p_i = \log^{-a} i$ for homophyly exponent $a > 4$ and for $d \geq 4$. Let the uniform threshold $\phi = O(\frac{1}{\log^b n})$ for $b = \frac{a}{2} - 2 - \epsilon$ for arbitrarily small $\epsilon > 0$.*

Then, with probability $1 - o(1)$ (over the construction of G), there is no initial set of polylogarithmic size that causes a cascading failure set of nonnegligible size. Precisely, we have that for any constant $c > 0$,

$$\Pr_{G \in_{\mathrm{R}} \mathcal{S}(n,a,d),\ G=(V,E)} \left[\forall S \subseteq V,\ |S| = \lceil \log^c n \rceil,\ |\mathrm{inf}^{\phi}_G(S)| = o(n) \right] = 1 - o(1),$$

where $\mathrm{inf}^{\phi}_G(S)$ is the infection set of S in G with uniform threshold ϕ.

By Theorem 6.9, if $a > 4$ and $d \geq 4$, then for $\phi = O(1/\sqrt{\log^b n})$, networks constructed by the security model $\mathcal{S}(n, a, d)$ are ϕ-secure. Here, ϕ is arbitrarily close to 0, i.e., $\phi = o(1)$. Therefore, by Definition 3.2, for $a > 4$ and $d \geq 4$, networks in $\mathcal{S}(n, a, d)$ are secure under the uniform threshold cascading failure model of attacks.

For the random threshold cascading failure model, each node v picks randomly, uniformly, and independently a number r from $1, 2, \ldots, d_v$, and defines its threshold $\phi(v) = \frac{r}{d_v}$, where d_v is the degree of v in G. Let $\mathrm{inf}^{\mathrm{R}}_G(S)$ be the infection set of attacks on S in G. We show that graphs generated by $\mathcal{S}(n, a, d)$ are secure.

Theorem 6.10. (Random threshold security theorem.) *Let $a > 6$ be the homophyly exponent, and $d \geq 4$. Suppose that G is a graph generated from $\mathcal{S}(n, a, d)$.*

Then, with probability $1 - o(1)$ (over the construction of G), there is no initial set of polylogarithmic size that causes a cascading failure set of nonnegligible size. Formally, we have that for any constant $c > 0$,

$$\Pr_{G \in_{\mathrm{R}} \mathcal{S}(n,a,d),\ G=(V,E)} \left[\forall S \subseteq V, |S| = \lceil \log^c n \rceil, |\mathrm{inf}^{\mathrm{R}}_G(S)| = o(n) \right] = 1 - o(1).$$

Theorems 6.9 and 6.10 show that for appropriately chosen parameters, networks constructed from the security model are provably secure for any attacks of small sizes under both uniform and random threshold cascading failure models. By Definitions 3.2, 3.1, 3.3, and by Theorems 6.9 and 6.10, the security model in Definition 5.1 is secure.

Finally, we show that the networks of the security model are not only provably secure, but they also satisfy the engineering requirement of communication networks.

Theorem 6.11. **(Expander core theorem.)** *For $a > 0$, $d \geq 3$, let G be a network of the security model with the given parameters, and let K be the induced subgraph of all the seed nodes of G in G. Then, for any constants $c > 0$ and $0 < \alpha < \min\{\frac{d-2}{2} - \frac{c+1}{4}, \frac{(d-2)\ln 2 - 1/e}{2+\ln 4(d-1)}\}$, with probability $1 - o(|K|^{-c})$, the conductance of K is $\Phi(K) \geq \frac{\alpha}{2d+\alpha}$.*

Theorem 6.11 indicates that, almost surely, the induced subgraph K of the seed nodes of a network G of the security model is an expander, that is, the conductance of K is $\Phi(K) \geq \beta$ for some constant β.

The remaining sections are devoted to proofs of Theorems 6.1, 6.4, 6.6, 6.8, 6.9, and 6.10. In Section 7, we prove Theorem 6.1. In Section 8, we prove Theorems 6.4, and 6.6. In Section 9, we establish Theorem 6.8. In Section 10, we prove Theorems 6.9, and 6.10 by using Theorems 6.1, 6.4, 6.6, and 6.8. In Section 11, we prove the expander core theorem, i.e., Theorem 6.11. In Section 12, we analyze the structural principles of security of networks and propose a protocol for secure computer network based on the security model of networks. In Section 13, we summarize the conclusions and discuss some future directions.

7. THE FUNDAMENTAL THEOREM OF THE SECURITY MODEL

In this section, we prove Theorem 6.1.

Before proving the theorem, we state the Chernoff bound, which will be frequently used in our proofs.

Lemma 7.1. **(Chernoff bound, [9].)** *Let $X_1, \ldots, X_n$ be independent random variables with $\Pr[X_i = 1] = p_i$ and $\Pr[X_i = 0] = 1 - p_i$. Denote the sum by $X = \sum_{i=1}^n X_i$ with expectation $E(X) = \sum_{i=1}^n p_i$. Then we have*

$$\Pr[X \leq E(X) - \lambda] \leq \exp\left(-\frac{\lambda^2}{2E(X)}\right),$$

$$\Pr[X \geq E(X) + \lambda] \leq \exp\left(-\frac{\lambda^2}{2(E(X) + \lambda/3)}\right).$$

Let G be a network constructed from the security model. We now prove Theorem 6.1. We will prove (1), (2), (3), and (4) of Theorem 6.1 in Subsections 7.1, 7.2, 7.3, and 7.4, respectively.

7.1. Basic Properties

In this subsection, we prove Theorem 6.1(1). It consists of two results, the first is the estimation of the number of seed nodes, and the second is the upper bound of sizes of the homochromatic sets.

Proof. **(Proof of (1) of Theorem 6.1)** We use $G[t]$ to denote the graph constructed at the end of time step t of the construction of G. Let $T_1 = \log^{a+1} n$ and C_t be the set of all colors appearing in $G[t]$.

For (i). It suffices to show that the size of C_t is bounded as desired. For this, we have

Lemma 7.2. *With probability* $1 - o(1)$, *for all* $t \geq T_1$, $\frac{t}{2\log^a t} \leq |C_t| \leq \frac{2t}{\log^a t}$.

Proof. The expectation of $|C_t|$ is

$$E[|C_t|] = 2 + \sum_{i=3}^{t} \frac{1}{\log^a i}.$$

Let t_0 be an integer such that $1/\log^a t_0 \geq 6a/\log^{a+1} t_0$, so for all $t \geq t_0$, $1/\log^t \geq 6a/\log^{a+1} t$. By the indefinite integral

$$\int \left(\frac{1}{\log^a x} - \frac{a}{\log^{a+1} x} \right) dx = \frac{x}{\log^a x} + C,$$

we know that

$$\begin{aligned}
\sum_{i=3}^{t} \frac{1}{\log^a i} &\leq 1 + \int_2^t \frac{1}{\log^a x} dx \\
&\leq \int_2^{t_0} \frac{1}{\log^a x} dx + \int_{t_0}^{t} \frac{1}{\log^a x} dx \\
&\leq \int_2^{t_0} \frac{1}{\log^a x} dx + \int_{t_0}^{t} \frac{6}{5} \left(\frac{1}{\log^a x} - \frac{a}{\log^{a+1} x} \right) dx \\
&= \int_2^{t_0} \frac{1}{\log^a x} dx + \frac{6}{5} \left(\frac{t}{\log^a t} - \frac{t_0}{\log^a t_0} \right).
\end{aligned}$$

Because $\int_2^{t_0} \frac{1}{\log^a x} dx - \frac{6}{5} \cdot \frac{t_0}{\log^a t_0}$ is finite, we have that, for large enough t,

$$\sum_{i=3}^{t} \frac{1}{\log^a i} \leq \frac{4t}{3\log^a t}.$$

Similarly, we can show that

$$\sum_{i=3}^{t} \frac{1}{\log^a i} \geq \frac{3t}{4\log^a t}.$$

By the Chernoff bound and the fact that $t \geq T_1 = \log^{a+1} n$, with probability $1 - exp(-\Omega(\frac{t}{\log^a t})) = 1 - o(n^{-1})$, we have $\frac{t}{2\log^a t} \leq |C_t| \leq \frac{2t}{\log^a t}$. By the union bound, such an inequality holds for all $t \geq T_1$ with probability $1 - o(n^{-1}) \cdot n = 1 - o(1)$. □

So, (i) follows from Lemma 7.2.

Lemma 7.2 depends on only the probability $p_i = 1/(\log i)^a$ with which the node created at time step i chooses a new color. It is a useful fact throughout the proofs, from which we define the following:

Definition 7.3. We define $\mathcal{E}$ to be the event that $|C_t|$ is bounded in the interval $\left[\frac{t}{2\log^a t}, \frac{2t}{\log^a t}\right]$.

By Lemma 7.2, almost surely, the event $\mathcal{E}$ holds for all $t \geq T_1$.

For (ii). We estimate the size of all the homochromatic sets.

Lemma 7.4. *With probability* $1 - o(1)$*, the following properties hold:*

(1) Every community has size bounded by $O(\log^{a+1} n)$*, and*
(2) For every $t \geq T_1$*, every community at the end of time step* t *has size bounded by* $O(\log^{a+1} t)$*.*

Proof. For (1). It suffices to show that with probability $1 - o(n^{-1})$, the homochromatic set of the first color κ has size $O(\log^{a+1} n)$.

We define an indicator random variable Y_t for the event that the vertex created at time t chooses color κ. We also define $\{Z_t\}$ to be the independent Bernoulli trails such that

$$\Pr[Z_t = 1] = \left(1 - \frac{1}{\log^a n}\right)\frac{2\log^a t}{t}.$$

Conditioned on the event $\mathcal{E}$, we know that $Y := \sum_{t=1}^{n} Y_t$ is stochastically dominated by $Z := \sum_{t=1}^{n} Z_t$. The latter has an expectation

$$E[Z] \leq \sum_{t=1}^{n} \frac{2\log^a t}{t} \leq 2\log^{a+1} n.$$

By the Chernoff bound,

$$\Pr[Z > 4\log^{a+1} n] \leq n^{-1}.$$

Therefore, with probability $1 - n^{-1}$, the size of S_κ is $Y \leq 4\log^{a+1} n$; (1) follows.

For (2). This follows from the proof of (1); (2) holds.

Lemma 7.4 follows. □

So, (ii) holds.

This proves (1) of Theorem 6.1.

7.2. Power Law and Holographic Law

In this subsection, we prove (2) of Theorem 6.1, consisting of power law of the induced subgraph of communities, of the degree distributions of the homochromatic sets, and of the whole network G.

Before proving the results, we first prove both a lower and an upper bound for the sizes of well-evolved communities.

Recall that $T_1 = \log^{a+1} n$. Let $T_2 = (1 - \delta_1)n$, for $\delta_1 = \frac{10}{\log^{a-1} n}$. We have

Lemma 7.5. *With probability* $1 - o(1)$*, both (1) and (2) below hold in* G*:*

(1) For a community created at a time step $\leq T_2$*, it has size at least* $\log n$*;*
(2) For a community created at a time step $> T_2$*, it has size at most* $30\log n$*.*

Proof. For (1). We need to prove only that, on the condition of event $\mathcal{E}$ in Definition 7.3, any homochromatic set S_κ created before time step $T_2 + 1$ has size at least $\log n$ with probability $1 - o(n^{-1})$.

For every $t > T_2$, let Y_t be the indicator random variable that the vertex, v say, created at time step t and that chooses old color κ. For $t > T_2$, let $\{Z_t\}$ be the independent Bernoulli trails such that

$$\Pr[Z_t = 1] = \left(1 - \frac{1}{\log^a (1-\delta_1)n}\right) \frac{\log^a t}{2t}. \tag{7.1}$$

Conditioned on the event $\mathcal{E}$, we know that $Y := \sum_{t \geq T_2+1}^{n} Y_t$ stochastically dominates $Z := \sum_{t \geq T_1+1}^{n} Z_t$, which has expectation

$$E[Z] \geq \sum_{t=T_2+1}^{n} \frac{\log^a t}{2t} \geq \frac{\delta_1}{2} \log^a (1-\delta_1)n \geq 4 \log n.$$

By the Chernoff bound,

$$\Pr[Z < \log n] \leq e^{-\frac{3^2 \log n}{2\times 4}} = n^{-\frac{9}{8}}.$$

Thus, with probability $1 - o(n^{-1})$, the size of S_κ is at least $\log n$.

For (2). The proof is similar to that of (1). We need to prove only that, on the condition of event $\mathcal{E}$, any homochromatic set S_κ created after T_2 has size at most $30 \log n$ with probability $1 - o(n^{-1})$. For $t > T_2$, we consider the Bernoulli random variables $\{Z_t\}$ defined by

$$\Pr[Z_t = 1] = \left(1 - \frac{1}{\log^a n}\right) \frac{2 \log^a t}{t}. \tag{7.2}$$

Note that

$$E[Z] \leq \sum_{t=T_2+1}^{n} \frac{2 \log^a t}{t} \leq \frac{2\delta_1}{1-\delta_1} \log^a n.$$

By a similar analysis to that in (1), we know that with probability $1 - o(n^{-1})$, the size of S_κ is at most $30 \log n$. □

The proof of Lemma 7.5 depends on both the probability $1 - p_i$ with which the newly created node chooses an old color, and the randomness and uniformity of the choice of the old color at time step i for all i's.

By Lemma 7.5, we know that each of the communities born before time step $T_2 + 1$ has expected size $\omega(1)$, and that all the communities born at time steps $\leq T_2$ account for $(1 - o(1))$ of all the communities. Therefore, we prove the power-law distribution only for the communities born at time steps $\leq T_2$.

For both (i) and (ii) of Theorem 6.1(2). Now we turn to prove two results:

(a) For each homochromatic set X, the degrees of nodes in X follow a power law, and
(b) For each homochromatic set X, the induced subgraph G_X of X follow a power law.

We prove both (a) and (b) together. We consider only the nontrivial homochromatic sets, i.e., the well-evolved communities, by ignoring the few most recently created communities.

By (4) of Definition 5.1, each community basically follows the classical PA model; we are able to give explicit expressions for the expected numbers of nodes of degree k for all k, for each of the homochromatic sets and for the induced subgraphs of the homochromatic sets.

In fact, as we will show following, the contribution to the degrees of a homochromatic set from the global edges is much smaller than that from the local edges of the homochromatic set. This is the key point to our proofs of the power law of almost all the communities.

We use X to denote a homochromatic set of a fixed color, κ say. Let T_0 be the time step at which X is created.

For positive integers s and k, we define $A_{s,k}$ to be the number of nodes of degree k in X when $|X|$ reaches s, $B_{s,k}$ to be the number of nodes of degree k in the induced subgraph of X when $|X|$ reaches s, and $g_{s,k}$ to be the number of global edges associated with the nodes in X of degree k in the induced subgraph of X when $|X|$ reaches s. By the initial condition of the definition, we have $A_{1,d} = 1$ and $A_{1,k} = 0$ for all $k > d$, and $B_{1,k} = 0$ for all k. Then we establish the recurrence formula for the expectations of both $A_{s,k}$ and $B_{s,k}$.

First, we define some notations associated with X and its size $|X|$:

- we use $T(s)$ (or T, for simplicity) to denote the time step at which the size of X becomes s,
- we use s_1 to denote the number of global edges connecting to X in the case that $|X| = s$.

We consider the time interval $(T(s-1), T(s))$. Then, the number of times that a global edge is created and linked to a node in X of degree k at some time step in the interval $(T(s-1), T(s))$ is expected to be $\Theta(\frac{1}{\log^a T} \cdot \frac{k \cdot A_{s,k}}{2dT} / \frac{\log^a T}{T}) = \Theta(\frac{k \cdot A_{s,k}}{\log^{2a} T})$. Denote $\Theta(\log^{2a} T)$ by s_2.

Then, for $s > 1$ and $k > d$, we have

$$E(A_{s,k}) = A_{s-1,k}\left(1 - \frac{kd}{2d(s-1)+s_1} - \frac{k}{s_2}\right) + A_{s-1,k-1} \cdot \left(\frac{(k-1)d}{2d(s-1)+s_1} + \frac{k-1}{s_2}\right) + O\left(\frac{1}{s^2}\right).$$

Taking expectations on both sides, we have

$$E(A_{s,k}) = E(A_{s-1,k})\left(1 - \left(\frac{1}{2(s-1)+s_1/d} - \frac{1}{s_2}\right)k\right) + E(A_{s-1,k-1})\left(\frac{1}{2(s-1)+s_1/d} + \frac{1}{s_2}\right)(k-1) + O\left(\frac{1}{s^2}\right). \tag{7.3}$$

If $k = d$, then

$$E(A_{s,d}) = E(A_{s-1,d})\left(1 - \left(\frac{1}{2(s-1)+s_1/d} - \frac{1}{s_2}\right)d\right) + 1 + O\left(\frac{1}{s^2}\right). \tag{7.4}$$

Similarly, for $s > 1$ and $k > d$,

$$E(B_{s,k}) = B_{s-1,k} - \frac{d \cdot (kB_{s-1,k} + g_{s-1,k})}{2d(s-1)+s_1} + \frac{d \cdot ((k-1)B_{s-1,k-1} + g_{s-1,k-1})}{2d(s-1)+t_1} + O\left(\frac{1}{s^2}\right).$$

Taking expectations on both sides, we have

$$\begin{aligned} E(B_{s,k}) &= E(B_{s-1,k})\left(1 - \frac{kd}{2d(s-1)+s_1}\right) + E(B_{s-1,k-1}) \cdot \frac{(k-1)d}{2d(s-1)+s_1} \\ &+ \frac{E(g_{s-1,k-1} - g_{s-1,k})}{2d(s-1)+s_1} + O(\frac{1}{s^2}). \end{aligned} \tag{7.5}$$

If $k = d$, then

$$\begin{aligned} E(B_{s,d}) &= B_{s-1,d} - \frac{d \cdot (dB_{s-1,d} + g_{s-1,d})}{2d(s-1)+s_1} + 1 + O\left(\frac{1}{s^2}\right) \\ &= B_{s-1,d}\left(1 - \frac{d}{2(s-1)+s_1/d}\right) + \left(1 - \frac{g_{s-1,d}}{2d(s-1)+s_1}\right), \end{aligned} \tag{7.6}$$

and

$$E(B_{s,d}) = E(B_{s-1,d})\left(1 - \frac{d}{2(s-1)+s_1/d}\right) + \left(1 - \frac{E(g_{s-1,d})}{2d(s-1)+s_1}\right).$$

To solve the recurrences, we invoke the following lemma.

Lemma 7.6. **(Lemma 3.1 [10].)** *Suppose that a sequence $\{a_s\}$ satisfies the recurrence relation*

$$a_{s+1} = \left(1 - \frac{b_s}{s+s_1}\right) a_s + c_s \quad \textit{for} \quad s \geq s_0,$$

where the sequences $\{b_s\}$, $\{c_s\}$ satisfy $\lim_{s\to\infty} b_s = b > 0$ and $\lim_{s\to\infty} c_s = c$ respectively. Then the limitation of $\frac{a_s}{s}$ exists and

$$\lim_{s\to\infty} \frac{a_s}{s} = \frac{c}{1+b}.$$

For the recurrence of $E(A_{s,k})$, by Lemma 7.5, as n goes to infinity, $t = \omega(1)$ also goes to infinity. By the definition of s_2, $s_2 = \Theta(\log^{2a} T) = \omega(s)$.

To deal with s_1, we give an upper bound for the expected volume of X at time T, denoted by V_T, as follows.

$$\begin{aligned} E(V_T) &\leq \sum_{i=2}^{T} \left[\left(1 - \frac{1}{\log^a i}\right) \cdot \frac{2d}{|C_i|} + \frac{1}{\log^a i} \cdot \frac{dV_{i-1}}{2di}\right] \\ &\leq \sum_{i=2}^{T} \frac{2d}{|C_i|} = O\left(\sum_{i=2}^{T} \frac{4d \log^a i}{i}\right) = O(\log^a T). \end{aligned}$$

So, it is easy to observe that $\frac{s_1}{t} = O\left(\frac{1}{\log^a T} \cdot \frac{V_T}{2dT} / \frac{\log^a T}{T}\right) = O\left(\frac{1}{\log^{a-1} T}\right)$ goes to zero as s approaches to infinity.

For the recurrence of $E(B_{s,k})$, we show that as s goes to infinity, both $\frac{E(g_{s-1,k-1} - g_{s-1,k})}{2d(s-1)+s_1}$ and $\frac{E(g_{s-1,d})}{2d(s-1)+s_1}$ approach to 0. Define $g_s = \sum_i g_{s,i}$ to be the total number of global edges associated to X when $|X|$ reaches s. We only have to show that $E(\frac{g_s}{s}) \to 0$ as $s \to \infty$.

Suppose that the seed node of X is created at time T_0.

$$E(g_s) = O\left(\sum_{i=T_0}^{T(s)} \frac{1}{\log^a i} \cdot \frac{V_i}{2di}\right) = O\left(\sum_{i=T_0}^{T(s)} \frac{\log i}{2di}\right) = O(\log^2 T(s) - \log^2 T_0).$$

Note that when we consider the size of X at time step $t > T_0$, we have

$$E(|X|) = \sum_{i=T_0}^{t}\left(1 - \frac{1}{\log^a i} \cdot \frac{1}{|C_i|}\right) = \Omega\left(\sum_{i=T_0}^{t} \frac{\log^a i}{2i}\right)$$
$$= \Omega\left(\int_{T_0}^{t} \frac{\log^a x}{2x} dx\right) = \Omega(\log^{a+1} t - \log^{a+1} T_0).$$

Thus, at time step $T(s)$, by the Chernoff bound, with probability $1 - o(1)$, $s = \Omega(\log^{a+1} T(s) - \log^{a+1} T_0)$. Therefore, $E(g_s) = o(s)$, that is, $E(\frac{g_s}{s}) \to 0$ as $s \to \infty$.

Then, we turn to consider the recurrences of $E(A_{s,k})$ and $E(B_{s,k})$. The terms s_1/d and $\frac{1}{s_2}$ in equalities (7.3) and (7.4) are comparatively negligible. The terms $\frac{E(g_{s-1,k-1} - g_{s-1,k})}{2d(s-1)+s_1}$ and $\frac{E(g_{s-1,d})}{2d(s-1)+s_1}$ in equalities (7.5) and (7.6), respectively, are also comparatively negligible. By Lemma 7.6, $\frac{E(A_{s,k})}{s}$ and $\frac{E(B_{s,k})}{s}$ must have the same limit as t goes to infinity. Next, we will give the proof only of the power-law distribution for $E(A_{s,k})$, which also holds for $E(B_{s,k})$.

Denote by $S_k = \lim_{t\to\infty} \frac{E(A_{s,k})}{s}$ for $k \geq d$. In the case of $k = d$, we apply Lemma 7.6 with $b_s = \frac{d}{2}$, $c_s = 1 + O(\frac{1}{s^2})$, $s_1 = -1$, and get

$$S_d = \lim_{s\to\infty} \frac{E(A_{s,d})}{t} = \frac{1}{1 + \frac{d}{2}} = \frac{2}{2 + d}.$$

For $k > d$, assume that we already have $S_{k-1} = \lim_{t\to\infty} \frac{E(A_{s,k-1})}{t}$. Applying Lemma 7.6 again with $b_s = \frac{k}{2}$, $c_s = \frac{E(A_{s-1,k-1})}{s-1} \cdot \frac{k-1}{2}$, $s_1 = -1$, we get

$$S_k = \lim_{t\to\infty} \frac{E(A_{s,k})}{s} = \frac{S_{k-1} \cdot \frac{k-1}{2}}{1 + \frac{k}{2}} = S_{k-1} \cdot \frac{k-1}{k+2}.$$

Thus, recurrently, we have

$$S_k = S_d \cdot \frac{(d+2)!(k-1)!}{(d-1)!(k+2)!} = \frac{2d(d+1)}{k(k+1)(k+2)}. \tag{7.7}$$

This implies

$$|E(A_{s,k}) - S_k \cdot s| = o(s),$$

and, thus,

$$E(A_{s,k}) = (1 + o(1))k^{-3}s.$$

Because $s = \omega(1)$ goes to infinity as $n \to \infty$, $E(A_{s,k}) \propto k^{-3}$. For the same reason, $E(B_{s,k}) \propto k^{-3}$. This proves (a) and (b), and also completes the proofs of both (i) and (ii) of Theorem 6.1(2).

For (iii). For the whole network, a key observation is that the union of several power-law distributions is also a power-law distribution if the powers are equal. We will give the same explicit expression of the expectation of the number of degree k nodes by combining those for the homochromatic sets, leading to a similar power-law distribution.

To prove the power-law degree distribution of the whole graph, we take the union of distributions of all homochromatic sets. We will show that, with overwhelming probability, almost all nodes belong to some large homochromatic sets, so that the role of small homochromatic sets is negligible.

Suppose that G has m homochromatic sets of size at least $\log n$. For $i = 1, \dots, m$, let M_i be the size of the ith homochromatic set and $N^{(i)}_{s,k}$ denote the number of nodes of degree k when the ith set has size s. For each i, we have

$$\lim_{n\to\infty} \frac{E(N^{(i)}_{M_i,k})}{M_i} = S_k.$$

Hence,

$$\lim_{n\to\infty} \frac{E(\sum_{i=1}^m N^{(i)}_{M_i,k})}{\sum_{i=1}^m M_i} = S_k.$$

Let M_0 denote the size of the union of all other homochromatic sets of size less than $\log n$, and $N^{(0)}_{s,k}$ denote the number of nodes of degree k in this union when it has size s. By Lemma 7.5, with probability $1 - o(1)$, all these sets are created after time T_2, and thus, $M_0 \le n - T_2 = \frac{10n}{\log^{a-1} n} = o(n)$.

Define $N_{t,k}$ to be the number of nodes of degree k in G_t, that is, the graph obtained after time step t. Then we have

$$\lim_{n\to\infty} \frac{E(N_{n,k})}{n} = \lim_{n\to\infty} \frac{E(\sum_{i=0}^m N^{(i)}_{M_i,k})}{\sum_{i=0}^m M_i}.$$

For M_0, we have that

$$\lim_{n\to\infty} \frac{M_0}{\sum_{i=1}^m M_i} = \lim_{n\to\infty} \frac{M_0}{n - M_0} = 0$$

and

$$\lim_{n\to\infty} \frac{E(N^{(0)}_{M_0,k})}{n} \le \lim_{n\to\infty} \frac{M_0}{n} = 0$$

hold with probability $1 - o(1)$. So,

$$\lim_{n\to\infty} \frac{E(N_{n,k})}{n} = \lim_{n\to\infty} \frac{E(\sum_{i=1}^m N^{(i)}_{M_i,k})}{\sum_{i=1}^m M_i} = S_k.$$

This implies

$$|E(N_{n,k}) - S_k \cdot n| = o(n),$$

and thus,

$$E(N_{n,k}) = (1 + o(1))k^{-3}n,$$

and $E(N_{n,k}) \propto k^{-3}$. (iii) follows.

For (iv). It follows from the proofs of (i)–(iii).

This completes the proof Theorem 6.1 (2). □

7.3. Small World Property

For Theorem 6.1(3). Now we turn to prove the properties of small diameters of each homochromatic set and the small-world phenomenon of the networks of the security model.

For (i). The diameter of the standard PA model is well-known [8,13], for which it has been shown that with probability $1 - O(\frac{1}{\log^2 n})$, a graph generated by the PA model, written $\mathcal{G}(n, d)$, has a diameter $O(\log n)$.

So, (i) follows immediately from Theorem 6.1(1)(ii).

For (ii). Now we prove the small-world phenomenon. We adjust the parameters in the proof of the PA model in [8] to get a weaker bound on diameters but a tighter probability. In so doing, we have the following lemma.

Lemma 7.7. *For any constant $a' > 2$, there is a constant K such that with probability $1 - \frac{1}{n^{a'+1}}$, a randomly constructed graph G from the PA model $\mathcal{P}(n, d)$ has a diameter $Kn^{1/(a'+1)}$.*

Proof. By a standard argument as that in the proof of the small-diameter property of networks of the preferential attachment [8]. □

Moreover, to estimate the distances among seed nodes, we recall a known conclusion on random recursive trees. A random recursive tree is constructed by stages as follows: at each stage, we create a new node, v say, and create an edge from v to a node chosen randomly and uniformly among all the nodes in the existing graph. A graph generated by this model is called a uniform recursive tree [33,34].

It was shown [39] that, with high probability, the height of a uniform recursive tree of size n is $O(\log n)$. We will use this result in our proofs.

Lemma 7.8. **([39].)** *With probability $1 - o(1)$, the height of a uniform recursive tree of size n is asymptotic to $e \log n$, where e is the natural logarithm.*

To estimate the average node-to-node distance of G, we assume that there are m homochromatic sets of size at most $\log n$. Choose a' in Lemma 7.7 to be the homophyly exponent a, and then we have a corresponding K.

Given a homochromatic set S, we say that S is *bad*, if the diameter of S is larger than $K|S|^{1/(a+1)}$.

We define an indicator X_S of the event that S is bad. Since $\log n \leq |S| = O(\log^{a+1} n)$, by Lemma 7.7, we have

$$\Pr[X_S = 1] \leq \frac{1}{\log^{a+1} n}.$$

By Lemma 7.2, the expected number of bad sets is at most $\frac{2n}{\log^a n} \cdot \frac{1}{\log^{a+1} n} = \frac{2n}{\log^{2a+1} n}$. By the Chernoff bound, with probability $1 - O(n^{-2})$, the number of bad sets is at most $\frac{3n}{\log^{2a+1} n}$. Thus the total number of nodes belonging to some bad set is at most $\frac{3n}{\log^a n}$. However, for any large set S that is not bad, its diameter is at most $K|S|^{1/(a+1)} = O(\log n)$.

Given two nodes u and v with distinct colors, suppose that c_0 and c_1 are the colors of u and v, respectively, that X and Y are the sets of nodes of colors c_0 and c_1, respectively, and that u_0 and v_0 are the seed nodes in X and Y, respectively. We consider a path from u to v as follows: (i) the first part is a path from u to u_0 within the induced subgraph of X, (ii) the second part is a path from u_0 to v_0 consisting of only global edges, and (iii) the third part is a path from v_0 to v, consisting of edges in the induced subgraph of Y. By the argument above, the number of the union of all bad homochromatic sets is bounded by $O(\frac{n}{\log^a n})$. By Definition 5.1, the giant connected component of all the seed nodes can be interpreted as a union of d uniform recursive trees. By Lemma 7.8, with probability $1 - o(1)$, there is a path from u_0 to v_0 in the induced subgraph of all seed nodes with length at most $O(\log n)$. Combining the three paths in (a), (b), and (c), we know that the average node-to-node distance in G is at most $O(\frac{\frac{2n^2}{\log^a n} \cdot \log^{a+1} n + n^2 \cdot \log n}{n^2}) = O(\log n)$. Theorem 6.1(3)(ii) follows.

For (iii). Suppose that G is a network constructed from the security model. We interpret G as a directed graph as follows: For an edge (u, v) in G, if u and v are created at time steps i, j, respectively, then for $i > j$, we identify the edge (u, v) as a directed edge (i, j).

We give an algorithm as follows: For any two nodes u, v in G,

1. Following the direction of time order in G (that is, an edge (x, y) means that y is created earlier than x) to find the seed nodes of the homochromatic sets of u and v, u_0 and v_0 say, respectively.
2. Take random walks from u_0 and v_0 in a directed uniform recursive tree of all the seed nodes created in 3(c) of Definition 5.1, until the two random walks cross.

By (i), Step 1 runs in time $O(\log \log n)$, by Lemma 7.8, Step 2 runs in time $O(\log n)$; (iii) follows.

This completes the proof of Theorem 6.1(3).

7.4. Small-Community Phenomenon

Before proving (4) of Theorem 6.1, we introduce some notations.

Let X be a homochromatic set, and x_0 be the seed node of X. We say that X is created at time step t, if the seed node x_0 of X is created at time step t. For $t \geq t_0$, we use $X[t]$ to denote the set of all nodes sharing the same color as that created at time step t_0 at the end of time step t. That is, we use $X[t]$ to denote a homochromatic set at the end of time step t.

We prove the small-community phenomenon stated in Theorem 6.1(4).

Intuitively speaking, we will show that the homochromatic sets created not too early or too late are good communities, with high probability. Then, the conclusion follows from the fact that the number of nodes in the remaining homochromatic sets takes up only a $o(1)$ fraction of the network.

We focus on the homochromatic sets created in time interval $[T_3, T_4]$, where $T_3 = \frac{n}{\log^{a+2} n}$, $T_4 = \left(1 - \frac{1}{\log^{(a-1)/2} n}\right) n$.

Given a homochromatic set S, we use t_S to denote the time step at which S is created.

Let S be a homochromatic set with $t_S \in [T_3, T_4]$, and let s be the seed node of S. For any $t \geq t_S$, we use $\partial(S)[t]$ to denote the set of edges from $S[t]$ to $\overline{S[t]}$, the complement of $S[t]$. By Definition 5.1, $\partial(S)[t]$ consists of two types of edges:

1. The edges from the seed node of $S[t]$ to earlier nodes, i.e., the edges of the form (t_S, j) for some $j < t_S$, and
2. The edges from the seed nodes created after time step t_S to nodes in $S[t]$

By Definition 5.1, the number of edges of type 1 is d.

We need to bound the number of edges of only the second type. For this, we first make an estimation on the total degrees of nodes in $S[t]$ at any given time step $t > t_S$.

For each $t \geq t_S$, we use $D(S)[t]$ to denote the total degree of nodes in $S[t]$ at the end of time step t of Definition 5.1. We have the following lemma.

Lemma 7.9. *For any homochromatic set S created at time $t_S \geq T_3$, $D(S)[n] = O(\log^{a+1} n)$ holds with probability $1 - o(1)$.*

Proof. We need only to show that for any $t \geq T_3$, if S is a homochromatic set created at time step t, then $D_n(S)[n] = O(\log^{a+1} n)$ holds with probability $1 - o(n^{-1})$. Without loss of generality, assume that S is created at time step $t_S = T_3$. The recurrence on $D(S)[t]$ can be written as

$$E[D(S)[t] \mid D(S)[t-1]] = D(S)[t-1] + \frac{1}{\log^a t}\left[\frac{D(S)[t-1]}{2d(t-1)} + (d-1)\cdot\frac{1}{|C_{t-1}|}\right] + \left(1 - \frac{1}{\log^a t}\right)\cdot\frac{2d}{|C_{t-1}|}.$$

We suppose again the event $\mathcal{E}$ that for all $t \geq T_1 = \log^{a+1} n$, $\frac{t}{2\log^a t} \leq |C_t| \leq \frac{2t}{\log^a t}$, which almost surely happens by Lemma 7.2. It holds also for $t \geq T_3$. On this condition,

$$\begin{aligned} &E[D(S)[t] \mid D(S)[t-1], \mathcal{E}] \\ &\leq D(S)[t-1]\left[1 + \frac{1}{\log^a t}\frac{1}{2d(t-1)}\right] + \frac{2d}{|C_{t-1}|} \\ &\leq D(S)[t-1]\left[1 + \frac{1}{\log^a t}\frac{1}{2d(t-1)}\right] + \frac{4d\log^a t}{t}. \end{aligned} \tag{7.8}$$

Then we use the submartingale concentration inequality (see [10], Chapter 2, for information on martingales) to show that $D(S)[t]$ is small, with high probability.

Because

$$\begin{aligned}&8d\log^{a+1}(t+1)-8d\left(1+\frac{1}{\log t}\frac{1}{2d(t-1)}\right)\cdot\log^{a+1}t\\ &\geq 8d\log^a t\left(\log\frac{t+1}{t}\right)-\frac{8d\log^a t}{2d(t-1)}\\ &\geq\frac{8d\log^a t}{t+1}-\frac{8d\log^a t}{2d(t-1)}\\ &\geq\frac{4d\log^a t}{t},\end{aligned}$$

applying it to inequality (7.8), we have

$$\begin{aligned}&E[D(S)[t]\mid D(S)[t-1],\mathcal{E}]-8d\log^{a+1}(t+1)\\ &\quad\leq(1+\frac{1}{\log t}\frac{1}{2d(t-1)})(D(S)[t-1]-8d\log^{a+1}t).\end{aligned}$$

For $t\geq T_3$, define $\theta_t=\Pi_{i=T_3+1}^{t}(1+\frac{1}{\log i}\frac{1}{2d(i-1)})$ and $X[t]=\frac{D(S)[t]-8d\log^{a+1}(t+1)}{\theta_t}$. Then,

$$E[X[t]\mid X[t-1],\mathcal{E}]\leq X[t-1].$$

Note that

$$X[t]-E[X[t]\mid X[t-1],\mathcal{E}]=\frac{D(S)[t]-E\left[D(S)[t]\mid D(S)[t-1],E\right]}{\theta_t}\leq 2d.$$

Because

$$D(S)[t]-D(S)[t-1]\leq 2d,$$

we have

$$\begin{aligned}\mathrm{Var}[X[t]\mid X[t-1],\mathcal{E}]&=E[(X[t]-E(X[t]|X[t-1],\mathcal{E}))^2]\\ &=\frac{1}{\theta_t^2}E[(D(S)[t]-E(D(S)[t]\mid D(S)[t-1],\mathcal{E}))^2]\\ &\leq\frac{1}{\theta_t^2}E[(D(S)[t]-D(S)[t-1])^2|D(S)[t-1],\mathcal{E}]\\ &\leq\frac{2d}{\theta_t^2}E[D(S)[t]-D(S)[t-1]\mid D(S)[t-1],\mathcal{E}]\\ &\leq\frac{2d}{\theta_t^2}\left[\frac{4d\log^a t}{t}+\frac{1}{\log^a t}\cdot\frac{D(S)[t-1]}{2d(t-1)}\right]\\ &=\frac{8d^2\log^a t}{t\theta_t^2}+\frac{1}{(t-1)\theta_t\log^a t}\cdot\frac{D(S)[t-1]}{\theta_t}\\ &\leq\frac{8d^2\log^a t}{t\theta_t^2}+\frac{8d\log^{a+1}t}{(t-1)\theta_t^2\log^a t}+\frac{X[t-1]}{(t-1)\theta_t\log^a t}\\ &\leq\frac{9d^2\log^a t}{t\theta_t^2}+\frac{X[t-1]}{2d(t-1)\theta_t\log^a t}.\end{aligned}$$

Note that θ_t can be bounded as

$$\theta_t \sim e^{\sum_{i=T_3+1}^{t} \frac{1}{2d(i-1)\log i}} \in \left[\left(\frac{t}{T_3}\right)^{\frac{1}{2d\log n}}, \left(\frac{t}{T_3}\right)^{\frac{1}{2d\log T_3}} \right].$$

Then,

$$\sum_{i=T_3+1}^{t} \frac{9d^2\log^a i}{i\theta_i^2} \le 9d^2\log^a n \int_{T_3}^{t} \frac{1}{i} \cdot \left(\frac{T_3}{i}\right)^{\frac{1}{d\log n}} di \le 9d^2\log^a n \cdot \log n = 9d^2\log^{a+1} n,$$

and

$$\sum_{i=T_3+1}^{t} \frac{1}{2d(i-1)\theta_i\log^a i} \le \frac{1}{d\log^a T_3} \int_{T_3}^{t} \frac{T_3^{\frac{1}{2d\log n}}}{i \cdot i^{\frac{1}{2d\log n}}} di \le \frac{\log n}{d\log^a T_3}.$$

Here, we can safely assume that $X[t]$ is nonnegative, which means that $D(S)[t] \ge 8\log^{a+1}(t+1)$, because otherwise, the conclusion follows immediately. Let $\lambda = 10\log^{a+1} n$. By the submartingale inequality ([10], Theorem 2.40),

$$\begin{aligned}\Pr[X[t] = \omega(\log^{a+1} n)] &\le \Pr[X[t] \ge X[T_3] + \lambda] \\ &\le \exp\left(-\frac{\lambda^2}{2(9d^3\log^{a+1} n + 10\log^{a+1} n + d\lambda/3)}\right) + O(n^{-2}) = O(n^{-2}).\end{aligned}$$

This implies that $D(S)[n] = O(\log^{a+1} n)$ holds, with probability $1 - O(n^{-2})$. □

Let S be a homochromatic set created at some time $t_S < T_4$. Let s be the seed node of S. We consider the edges from seed nodes created after time step t_S to nodes in S. For $t > t_S$, if a seed node, v say, is created at time step t, then there are two types of edges from v to nodes in $S[t-1]$, they are:

1. (First-type edges) An edge (v, u) for some $u \in S[t-1]$ created by Step 3(b) of Definition 5.1.

   We call these edges the *first-type edges*.
2. (Second-type edges) Some edges (v, s) for the seed node $s \in S[t-1]$ created by Step 3(c) of Definition 5.1.

   We call these edges the *second-type edges*.

We will bound the numbers of these two types of edges, respectively.

By a similar proof to that in Lemma 7.5(1), we are able to show that, with probability $1 - o(1)$, $S = S[n]$ has a size $\Omega(\log^{\frac{a+1}{2}} n)$, and so a volume $\Omega(\log^{\frac{a+1}{2}} n)$. We suppose the event, denoted by $\mathcal{F}$, that for any $t \ge T_S$, $D(S)[t] = O(\log^{a+1} n)$, which holds with probability $1 - o(1)$ by Lemma 7.9. For each $t \ge T_S$, we define a 0, 1 random indicator variable X_t, which indicates the event that the first-type edge connects to S at time t and satisfies

$$\Pr[X_t = 1|\mathcal{F}] = \frac{1}{\log^a t}\frac{D(S)[t-1]}{2d(t-1)} \le \frac{\log^{1+\epsilon} n}{2d(t-1)},$$

for arbitrarily small positive ϵ, i.e., $0 < \epsilon < \frac{a-1}{4}$. Then,

$$E\left[\sum_{t=t_S}^{n} X_t\right] \leq \log^{1+\epsilon} n \sum_{t=t_S}^{n} \frac{1}{2d(t-1)} \leq (\log^{1+\epsilon} n)(\log\log n).$$

By the Chernoff bound,

$$\Pr\left[\sum_{t=t_S}^{n} X_t \geq 2(\log^{1+\epsilon} n)(\log\log n)\right] \leq n^{-2}.$$

That is, with probability at least $1 - n^{-2}$, the total number of edges of the first type is upper bounded by $2(\log^{1+\epsilon} n)(\log\log n)$.

For the second type of edges, conditioned on the event $\mathcal{E}$, this number is expected to be at most

$$\sum_{t=T_3}^{n} \frac{1}{\log^a t} \cdot \frac{1}{|C_t|} \cdot (d-1) \leq O\left(\sum_{t=T_3}^{n} \frac{1}{\log^a t} \cdot \frac{2\log^a t}{t}\right) = O(\log\log n).$$

So, by the Chernoff bound, with probability $1 - o(1)$, the number of edges of the second type is upper bounded by $O(\log n)$.

Hence, with probability $1 - o(1)$, the conductance of S is

$$\Phi(S) = O\left(\frac{2(\log^{1+\epsilon} n)(\log\log n) + \log n}{\log^{(a+1)/2} n}\right) \leq O\left(\log^{-\frac{a-1}{4}} n\right) \leq O\left(|S|^{-\frac{a-1}{4(a+1)}}\right).$$

The total number of nodes belonging to the homochromatic sets that appear before time T_3 or after time T_4 is at most $\log^{a+1} n \cdot \frac{n}{\log^{a+2} n} + \frac{n}{\log^{(a-1)/2} n} = o(n)$ for any constant $a > 1$. Therefore, $1 - o(1)$ fraction of nodes of G belongs to a subset W of nodes, which has a size bounded by $O(\log^{a+1} n)$ and a conductance bounded by $O(|W|^{-\frac{a-1}{4(a+1)}})$. This proves Theorem 6.1(4).

This completes the proof of Theorem 6.1.

8. COMBINATORIAL PRINCIPLES

Theorem 6.1 provides the necessary structural properties for proving Theorems 6.9, and 6.10. In this section, we prove the necessary probabilistic and combinatorial principles for the proofs of the security theorems, that is, Theorem 6.4, and Theorem 6.6.

8.1. Degree Priority Theorem

First, we establish a basic property of the expectation of the degrees of nodes of a network generated by the PA model.

Lemma 8.1. *Suppose that G is a network generated from the PA model. Let v_i be the ith vertex in G. Then we have that the degree of v_i is expected to be $\sqrt{\frac{n}{i}} \cdot d$.*

Proof. Let s_i be the expected degree of v_i. Fix i, and for $j \geq i$, let $a_i(j)$ be the expected degree contributed by v_j to v_i and $T_i(j)$ be the expected degree of v_i at the end of step j.

So for each i, $a_i(i) = T_i(i) = d$, $T_i(n) = s_i$, and $T_i(j) = \sum_{k=i}^{j} a_i(k)$. Note that the volume of the whole graph at step j is $2dj$. For $j \geq i$, $a_i(j+1) = \frac{T_i(j)}{2dj} \cdot d = \frac{T_i(j)}{2j}$, and hence, $T_i(j+1) = T_i(j) + \frac{T_i(j)}{2j}$. By this recurrence equation, we have

$$T_i(n) = \prod_{j=i}^{n-1} \left(1 + \frac{1}{2j}\right) \cdot T_i(i).$$

Define a function $f(m) = \prod_{j=1}^{m-1}(1 + \frac{1}{2j})$. So, $T_i(n) = \frac{f(n)}{f(i)} \cdot d$. Since $f(n) = \frac{(2n-1)!}{2^{2(n-1)}[(n-1)!]^2}$, by the Stirling formula, when n is large enough, $f(n) = \frac{2}{\sqrt{\pi}} \cdot \sqrt{n}$. Thus, $s_i = T_i(n) = \sqrt{\frac{n}{i}} \cdot d$. □

We turn to the proof of Theorem 6.4.

Proof. (Proof of Theorem 6.4) For (1). To bound the expected length of degrees for all nodes, we consider the following two cases.

Case 1. v is a nondeed node created at step t_0.

For every $t > t_0$, if a seed node u is created at step t, then the probability that an edge between v and u is created and is $\frac{d(v)[t]}{2dt}$, where $d(v)[t]$ is the degree of v in the graph at the beginning of step t. Let $X[t]$ be set of the the natural community of v at the end of step t. By Theorem 6.1(1)(ii), the size of $X[t]$ is $|X| = O(\log^{a+1} t)$. By Lemma 8.1, for every $t > t_0$, $d(v)[t] = O(\log^{\frac{a+1}{2}} t)$. By the construction of G in Definition 5.1, at step $t > t_0$, the probability that the newly born node is a seed, is $\frac{1}{\log^a n}$, so the expectation of the increment of the length of v at step t is $\frac{d(v)[t]}{2dt} \cdot \frac{1}{\log^a t}$, which is $O(\frac{1}{t})$ if the affinity exponent $a \geq 1$.

Therefore, for $a \geq 1$, the expected number of seed nodes created after time step t_0 and linked to v by step 3(b) of Definition 5.1 is at most

$$E[l(v)] = O\left(\sum_{t=t_0}^{n} \frac{1}{t}\right) = O(\log n).$$

Case 2. v is a seed node created at time t_0.

By the proof in Case 1, the length of v contributed by step 3(b) of Definition 5.1 is $O(\log n)$.

By Lemma 7.2, for each t, $|C_t|$ is expected to be $\Theta(\frac{t}{\log^a t})$. Thus, the expected number of seed nodes created after time t_0 and linked to v by step 3(c) of Definition 5.1 is at most $d \cdot \frac{1}{\log^a t} \cdot \frac{1}{|C_t|} = O(\frac{1}{t})$. This shows that

$$E[l(v)] = O\left(\sum_{t=1}^{n} \frac{1}{t}\right) = O(\log n).$$

Therefore, the expectation of the length of v is $O(\log n)$.

(1) follows.

For (2), (3), and (4). We prove (2)–(4) together by considering two cases:

Case 1. v is a nonseed node.

Suppose that v is created at time step t_0. We use $D(v)$ to denote the degree of v contributed by nodes of the same color as v, and $F(v)$ to denote the maximal degree of

v contributed by nodes that share the same color other than the color of v. By (4) of Definition 5.1, $D(v)[t_0] = d$, and $F(v)[t_0] = 0$.

For $t+1 > t_0$, let u be the node created at time step $t+1$. If u is a seed node, then by (3) of Definition 5.1, we have that $D(v)[t+1] = D(v)[t]$, and $F(v)[t+1] \leq \max\{F(v)[t], 1\}$. If u is a nonseed node, then either u has the same color as that of v, or u chooses an old color different from that of v; in either case, we have that $D(v)[t+1] \geq D(v)[t]$ and $F(v)[t+1] = F(v)[t]$.

Therefore, we have that the first degree of v, $d_1(v)$ is always contributed by the neighbors of v that share the same color as v, that is, $D(v)$, and that the second degree $d_2(v) \leq 1$.

Case 2. v is a seed node.

Let v be a node created at time step t_0. We use $F(v)[t]$ to denote the largest number of homochromatic neighbors having different color from v at the end of time step t.

By step (3) of Definition 5.1, $F(v)[t_0] \leq d$. For every $t \geq t_0$, we consider time step $t+1$. Let u be the node created at time step $t+1$. If u is a seed node, then by (3) of Definition 5.1, we have that $F(v)[t+1] \leq \max\{F(v)[t], d\}$. If u is a nonseed node, then by (4) of Definition 5.1, $F(v)[t+1] = F(v)[t]$.

Therefore, we have that $F(v)[n] \leq d = O(1)$.

Next we consider the degree of v contributed by the neighbors of the same color as v. Note that a seed node has a degree at least d contributed by local edges, unless the homochromatic set of the seed node is too small. This kind of seed node is likely to be created too late. We choose an appropriate time stamp T and show that there are only a negligible number of seed nodes born after T, and all the seed nodes born before $T+1$ are contained in homochromatic sets of nonnegligible size and, thus, have a large degree contributed by local edges.

Here we choose the time step $T = T_4$, defined in Subsection 7.4.

By the proof of Lemma 7.5, the homochromatic sets created at time step $\leq T_4$ has size at least $\Omega(\log^{\frac{a+1}{2}} n)$, with probability $1 - o(1)$. The next lemma guarantees that a seed node of a homochromatic set of size $\Omega(\log^{\frac{a+1}{2}} n)$ has degree $\Omega(\log^{\frac{a+1}{4}} n)$ contributed by local edges.

By Definition 5.1(4), the induced subgraph of a homochromatic set basically follows the PA scheme, so it suffices to prove a result for networks of the PA model.

So by step (3) of Definition 5.1, and by Lemma 8.1, with probability $1 - o(1)$, a homochromatic set of size at least $\Omega(\log^{\frac{a+1}{2}} n)$ has a seed node of degree at least $\Omega(\log^{\frac{a+1}{4}} n)$ contributed by local edges. So, the seed nodes created at time step $\leq T_4$ have their first degrees contributed by local edges with probability $1 - o(1)$.

By the proof in Subsection 7.4, the number of seed nodes created after time step T_4 is negligible.

Therefore, with probability $1 - o(1)$, a randomly picked seed node has its first degree contributed by its neighbors sharing the same color as the seed node.

All (2), (3), and (4) follow from Cases 1 and 2.

This completes the proof of Theorem 6.4. □

8.2. Inclusion-Exclusion Theorem

In this subsection, we prove Theorem 6.6. At first, we give a basic definition of communities, targeted communities, and infected communities.

Definition 8.2. Let G be a network constructed from the security model.

(1) A community of G is the induced subgraph of a homochromatic set of G.
(2) We say that a community, G_X say, is created at time step t, if the seed node of G_X is created at time step t.
(3) We say that a community, G_X say, is targeted, if there is a node in X that is targeted by an attack, and nontargeted, otherwise.
(4) We say that a community G_X is infected, if there is a node in X that has been either targeted or infected, and noninfected, otherwise.

Proof. **(Proof of Theorem 6.6)** For (1). We consider two cases:

For (i). The infection of G_Y from a nonseed node x_1 in G_X.

By Definition 5.1, there are no edges between nonseed nodes in G_X and nonseed nodes in G_Y, and there is no edge between the seed node of G_X and nonseed nodes in G_Y.

Therefore, there is no injury from G_X to any nonseed node in G_Y. Hence, the only possible node in G_Y that might be injured by G_X is the seed node y_0 of G_Y. (i) follows.

For (ii). The injury of the seed node in G_Y from G_X.

By Theorem 6.4, the number of neighbors of the seed node y_0 (of G_Y) in G_X is less than or equal to the second degree of y_0, which is at most a constant.

For (2). Suppose that x_1 and y_1 are nonseed nodes in X and Y, respectively.

For (i). The injury of G_Y from the nonseed node x_1.

This fails to occur because at the stage at which x_1 is created, it links to nodes only in G_X.

For (ii). The injury of the seed node y_0 of G_Y from the whole community G_X.

In this subcase, the possible neighbors of y_0 in G_X is only the seed x_0 of G_X, and y_0 is a seed node of G_Y. Therefore, the injury of y_0 from G_X is bounded by 1.

For (iii). The injury of a nonseed node y_1 from G_X.

The same as that in (i) and (ii), the only possible neighbor of y_1 in G_X is the seed node x_0 of G_X. In this case, by Definition 5.1, the only possibility that there is a link between x_0 and and a nonseed node y of G_Y is that y is the unique node chosen by the PA scheme in step 3(b) of Definition 5.1 at the time step at which x_0 is created.

Theorem 6.6(3) and (4) follow from (1) and (2).

This completes the proof of Theorem 6.6. □

9. INFECTION PRIORITY TREE PRINCIPLE

In this section, we define the infection priority tree, and prove the infection priority tree principle, i.e., Theorem 6.8.

At first, we introduce a notation of *injury* between nodes and communities.

Definition 9.1. **(Injury.)** Let x and y be two nodes, and G_X and G_Y be two natural communities.

(1) We define the injury of x from y to be the number of edges between x and y.
(2) We define the injury of x from G_Y to be the number of edges between x and the nodes in G_Y.

(3) We define the injury of G_X from G_Y to be the number of edges between nodes in G_X and the nodes in G_Y.

Then, we have that

Lemma 9.2. (Injury Lemma.) *For any communities G_X and G_Y, the injury of G_Y from the whole community G_X satisfies:*

1. *For the seed node y_0 of G_Y, the injury of y_0 from G_X is bounded by $O(1)$.*
2. *For a nonseed node $y \in Y$, G_X injures y, only if the following occurs:*
   - *y is injured only by the seed node x_0 of G_X,*
   - *y is created before the creation of the seed x_0 of G_X, and*
   - *At the time step at which x_0 is created, 3(b) of Definition 5.1 occurs, which creates an edge (x_0, y).*

Proof. By Theorem 6.6. □

By Theorem 6.1(2)(i), every community has size bounded by $O(\log^{a+1} n)$, it is harmless even if a targeted or infected node causes the infection of the whole community of the node. Therefore, we consider only the infections among different communities. By Lemma 9.2, we only consider two types of injuries among two communities.

Definition 9.3. (Injury Type.) We define:

1. (First type) The first type of injury is the injury of a seed node.
2. (Second type) The second type is an injury following an edge created by 3(b) of Definition 5.1.

To deal with the first type of injury, we introduce the notion of strong communities.

Definition 9.4. Given a homochromatic set X, suppose that x_0 is the seed node of X and that G_X is the community induced by X.

We say that G_X is a strong community if the seed node $x_0 \in X$ will never be infected, unless there is a node $x \in X$ that has already been infected. Precisely, X is strong if $\frac{d'(x_0)}{d(x_0)} < \phi(x_0)$, where $d'(x_0)$ is the number of edges from x_0 to nodes outside X, and $d(x_0)$ is the degree of x_0 in G. If X is not strong, we say that G_X is a vulnerable community.

By Theorem 6.4, for every seed node x_0 of a community G_X, the length of degrees of x_0 is bounded by $O(\log n)$, and the second degree of x_0 is bounded by $O(1)$, therefore the injury of the seed node x_0 from the collection of all the communities other than G_X itself can be bounded by $O(\log n)$. Notice that this upper bound of $O(\log n)$ is rather loose, and more importantly that the upper bound $O(\log n)$ is independent of the homophyly exponent a. However, by Theorem 6.4, the first degree of a seed node is lower bounded by $\Omega(\log^{\frac{a+1}{4}})$, which is large if the homophyly exponent a is appropriately large. This allows us to show that almost surely, for any set of attacks of polylogarithmic sizes, there is a huge number of strong communities.

By Lemma 9.2, the injury among strong communities is the second type. The infections among the strong communities is determined by the infection priority tree T of G by modulo the small communities from the network.

Definition 9.5. **(Infection of a strong community from a strong community.)** Suppose that X and Y are two natural communities. Let x_0 and y_0 be the seed nodes of X and Y, respectively. Let $\phi(x_0)$ and $\phi(y_0)$ be the thresholds of x_0 and y_0, respectively. Suppose that both X and Y are strong. We say that X is infected from Y, if there is an edge from y_0 to a nonseed node $x \in X$.

This means that the seed node $x_0 \in X$ cannot be infected by the nodes in Y. Hence, the only way that a node in X is infected by Y is that some nonseed node $x \in X$ is infected by Y, which occurs only through an edge between the seed node y_0 of Y to the nonseed node $x \in X$.

By the infection priority tree T in Definition 6.7, and by Definition 9.5, we have that

Lemma 9.6. *Any infection from a strong community to a strong community must be triggered by a directed edge in the infection priority tree T.*

Proof. By Definition 6.7, Definitions 9.4, 9.5, and Theorem 6.6. □

Lemma 9.6 shows that the cascading behavior in the infection priority tree T is always directed from a seed node to an old nonseed node created by 3(b) of Definition 5.1.

Now the key to our proofs is that the cascading procedure in T must terminate shortly, that is, after $O(\log n)$ many steps.

Proof. (Proof of Theorem 6.8) By Definition 5.1 and Definition 6.7, T can be regarded as a graph constructed by a PA scheme with $d = 1$ such that whenever a new node is created, it links to a node chosen with probability proportional to the weights of nodes while the weights of nodes are increasing uniformly and randomly. Precisely, we restate the construction of T as follows:

(i) Take H_2 to be a graph with two nodes $1, 2$, one directed edge $(2, 1)$ such that each node has a weight $w(i) = d$ for $i = 1, 2$.

For $i + 1 > 2$, let $p_i = 1/(\log i)^a$, and let H_i be the graph constructed at the end of time step i.

(ii) With probability p_i, we create a new node, v say, in which case,
   (a) let u_0 be a node chosen with probability proportional to the weights of nodes in H_i; create a directed edge (v, u_0),
   (b) let $u_1, u_2, \ldots, u_{d-1}$ be nodes chosen randomly and uniformly in H_i,
   (c) for each $j = 0, 1, \ldots, d - 1$, set $w(u) \leftarrow$ old $w(u) + 1$, and
   (d) set $w(v)[i + 1] = d$.

(iii) Otherwise, then choose randomly and uniformly a node, u say, in H_i, set $w(u)[i+1] = w(u)[i] + 2d$.

Then T is the directed graph obtained from H by ignoring the weights of nodes.

For (1). Clearly, it is true that T is a tree, because whenever one new node is created, there is only one new edge added, and the graph is connected; (1) holds.

For (2). By definition of T, the height of T is between a graph of the preferential attachment model with $d = 1$ and a uniform recursive tree of the same number of nodes. By Lemma 7.8, with probability $1 - o(1)$, a uniform recursive tree of nodes n has height bounded by $O(\log n)$. By this construction, T has height stochastically dominated by that of a uniform recursive tree of the same number of nodes. Therefore, with probability $1 - o(1)$, the height of T is bounded by $O(\log n)$; (2) holds. □

By Lemma 9.6 and Theorem 6.8, T exactly captures the cascading behaviors among strong communities, which is the key to our proofs.

Now we know that the proofs of both Theorem 6.9 and Theorem 6.10 consist of the following steps:

(1) To prove that for any attack of polylogarithmic size, almost surely, there is a huge number of strong communities.
(2) Any infection among the strong communities must be triggered by an edge in the infection priority tree of G, which goes at most $O(\log n)$ many steps, by Theorem 6.8.

Step (2) has been guaranteed by Lemma 9.6 and Theorem 6.8. So, the main issue for the proofs of Theorem 6.9 and Theorem 6.10 is actually Step (1) above, which will be given in Subsections 10.1 and 10.2.

10. SECURITY THEOREMS OF THE SECURITY MODEL

In this section, we will prove the security theorems of the security model, i.e., Theorems 6.9 and 6.10, by applying the fundamental theorem, i.e., Theorem 6.1, the probabilistic and combinatorial principles in Theorems 6.4, and 6.6, and the infection priority tree principle in Theorem 6.8.

10.1. Uniform Threshold Security Theorem

In this subsection, we prove Theorem 6.9.

Let G be a network constructed by the security model. Consider a deliberate attack by targeting an initial set S of size poly($\log n$). Note that the size of S, poly($\log n$), is much smaller than the number of communities, i.e., $\Theta(n/\log^a n)$, by (1)(i) of Theorem 6.1.

Proof. (Proof of Theorem 6.9) Set time $T_0 = (1 - \delta)n$, where $\delta = \log^{-b_0} n$, where b_0 will be determined later. We will show that with high probability, all the communities created before time step T_0 are both large and strong.

Lemma 10.1. *Let $2 < b_1 < a - b_0$. Then, with probability $1 - o(1)$, every homochromatic set created before time step T_0 has a size $\Omega(\log^{b_1} n)$.*

Proof. It is sufficient to show that, with probability $1 - n^{-1}$, for every homochromatic set S_κ created before T_0, S_κ has a size $\Omega(\log^{b_1} n)$.

Suppose that S_κ is the set with color κ, and that it is created at time step $t_0 \leq T_0$ for some t_0. For any $t \geq t_0$, define an indicator random variable Y_t to be the event that the node created at time step t chooses color κ.

Define $\{Z_t\}$ to be the independent Bernoulli trails such that

$$\Pr[Z_t = 1] = \left(1 - \frac{1}{\log^a (1-\delta)n}\right) \frac{\log^a t}{2t}.$$

Conditioned on the event $\mathcal{E}$ in Definition 7.3, we have that random variable $Y := \sum_{t=t'}^{n} Y_t$ stochastically dominates $Z := \sum_{t=t'}^{n} Z_t$ for any $t' \leq T_0$.

By definition, Z has an expectation

$$E[Z] \geq \left(1 - \frac{1}{\log^a (1-\delta)n}\right) \sum_{t=T_0+1}^{n} \frac{\log^a t}{2t} \geq \frac{\delta n}{2n} \log^a (1-\delta)n = \Omega(\log^{a-b_0} n).$$

Because $2 < b_1 < a - b_0$, by the Chernoff bound,

$$\Pr\left[Z = O(\log^{b_1} n)\right] \leq n^{-1}.$$

Therefore, with probability $1 - n^{-1}$, the size of S_κ is at least $Y = \Omega(\log^{b_1} n)$. □

Second, we show that every seed node created before T_0 probably has a large degree.

Lemma 10.2. *With probability $1 - o(1)$, every seed node created before time step T_0 has degree at least $\Omega(\log^{b_1/2} n)$.*

Proof. Let v be a seed node created at a time step $\leq T_0$. Suppose that v has color κ. Let S be the set of all nodes sharing color κ. Then the community G_S is the induced subgraph of S in G. The degree of the seed node v in G is contributed by both local edges and global edges. By the construction, G_S truthfully follows a power law, by Lemma 8.1, the degree of v contributed by local edges is expected to be at least $\sqrt{|S|}$. By Lemma 10.1, with probability $1 - o(1)$, each S has a size $\Omega(\log^{b_1} n)$. The degree of v has an expected degree of at least $\Omega(\log^{b_1/2} n)$. Because $b_1 > 2$, by the Chernoff bound, with probability $1 - o(n^{-1})$, v's degree is at least $\Omega(\log^{b_1/2} n)$. The lemma follows immediately by the union bound. □

Now we are able to estimate the number of strong communities.

Lemma 10.3. *Let $b_0 = 2 + \epsilon$ and $b_1 = a - b_0 - \frac{\epsilon}{2}$, where ϵ is that defined in Theorem 6.9. With probability $1 - o(1)$, all the communities created before time T_0 are strong.*

Proof. By Theorem 6.4, the length of degrees of a seed node is bounded by $O(\log n)$, and the second degree of a seed node is bounded by $O(1)$. By Chernoff bound, we have that, with probability $1 - o(1)$, for every seed node v, the degree of v contributed by global edges is bounded by $O(\log n)$. By Lemma 10.2, almost surely, for each seed node v, the fraction of v's degree contributed by global edges is less than or equal to $O(\log^{1-b_1/2} n)$. Recall that the threshold parameter $\phi = \Omega\left(\frac{1}{\log^b n}\right)$ for $b = \frac{a}{2} - 2 - \epsilon$ for arbitrary $\epsilon > 0$. By the choices of b_0 and b_1, $1 - b_1/2 = -\left(\frac{a}{2} - 2 - \frac{3\epsilon}{4}\right) < -b$. The lemma follows. □

For the total number of vulnerable communities, we have

Lemma 10.4. *Let $b_2 = a + b_0$. With probability $1 - o(1)$, the number of vulnerable communities is at most $\frac{2n}{\log^{b_2} n}$.*

Proof. By Lemma 10.3, we need to bound only the number of communities created after time step T_0. Because at time step t, a new color is created with probability $p_t = \log^{-a} t$, the number of colors created after time step T_0, denoted by N_{vul} is expected to be

$$E[N_{\text{vul}}] = \sum_{t=T_0+1}^{n} \frac{1}{\log^a t}.$$

When n is large enough, by a simple integral computation, $E[N_{\text{vul}}]$ is upper bounded by $\frac{3\delta n}{2\log^a n}$. By the Chernoff bound, with probability $1 - o(1)$, N_{vul} is at most $\frac{2\delta n}{\log^a n} = \frac{2n}{\log^{b_2} n}$. The lemma follows. □

Now we are ready for the proof of Theorem 6.9.

Suppose that S is the initially targeted set of size $\lceil \log^c n \rceil$. Choose $b_0 = 2 + \epsilon$, $b_1 = a - b_0 - \frac{\epsilon}{2}$ and $b_2 = a + b_0$.

By Lemma 10.4, with probability $1 - o(1)$, the number of vulnerable communities is at most $\frac{2n}{\log^{b_2} n}$. By Theorem 6.8, the height of infection priority tree T is $h = O(\log n)$. By Lemma 9.6, infections among strong communities must be triggered by an edge in the infection priority tree T. Therefore, the number of infected communities by attacks on S is at most

$$\left(|S| + \frac{2n}{\log^{b_2} n}\right) \cdot h = O\left(\left(\lceil \log^c n \rceil + \frac{2n}{\log^{b_2} n}\right) \cdot \log n\right).$$

By Theorem 6.1(1), with probability $1 - o(1)$, the largest community has a size $O(\log^{a+1} n)$. So, the number of infected nodes in G by attacks on S is at most

$$O\left(\left(\lceil \log^c n \rceil + \frac{2n}{\log^{b_2} n}\right) \cdot \log n \cdot \log^{a+1} n\right) = o(n).$$

This completes the proof of Theorem 6.9.

□

The proof of Theorem 6.9 is essentially a methodology of community analysis of networks of the security model. The key ideas of the methodology are those in Theorems 6.1, 6.4, and Theorem 6.6, Definition 9.4, Definition 6.7, Lemma 9.6, and Theorem 6.8.

The method allows us to divide all the communities into two classes, the first is the strong communities, and the second is the vulnerable ones. The two types of communities are distinguished by a time step T_0. This time stamp T_0 is determined by both parameter δ, and essentially by the power b. Then we show that communities created before time step T_0 are strong, and that the number of communities created after time step T_0 is small.

Theorem 6.9 shows that the power-law distribution in Theorem 6.1 is never an obstacle for security of networks. Our proof of the security theorem show that the community structure of the networks isolates the vulnerable nodes in a large number of small communities, that the homogeneity and randomness among the seed nodes or “hubs" guarantee that most communities are strong, and that the infection priority tree ensures that the cascading procedure among strong communities cannot be too long.

10.2. Random Threshold Security Theorem

In this subsection, we prove Theorem 6.10. The proof has the same framework as before. By Lemma 9.6, Theorem 6.8, infections among strong communities must be triggered by edges in the infection priority tree T, and infections in T are directed and terminate by $O(\log n)$ many steps.

Therefore, the only issue is to prove that the number of vulnerable communities is small.

Proof. Let $T_0 = (1 - \delta)n$, where $\delta = 100 \log^{-b_0} n$ and b_0 to be determined later. Let $T_0' = n/100$.

By a similar proof to that of Lemma 10.1, for every $b_1 \in (1, a - b_0]$, we have that, with probability $1 - o(1)$, the following hold:

- Every community created at a time step $t \leq T_0'$ has a size $\Omega(\log^a n)$, and
- Every community created at a time step $t \in [T_0', T_0]$ has a size $\Omega(\log^{b_1} n)$.

By the proof of Lemma 10.2, we have that, with probability $1 - o(1)$,

1. A seed node created at a time step $t \leq T_0'$ has degree $\Omega(\log^{a/2} n)$, and
2. A seed node created at a time step $t \in [T_0', T_0]$ has degree $\Omega(\log^{b_1/2} n)$.

Then, we show that the number of vulnerable communities created before time step T_0 is small.

Lemma 10.5. *Let $b_0 = \frac{a}{2} - 1$ and $b_1 = \frac{a}{2} + 1$. With probability $1 - o(1)$, there are only $O\left(\frac{n}{\log^{a+(b_1/2)} n}\right)$ communities created before time step T_0 that are vulnerable.*

Proof. By the Chernoff bound, with probability $1 - o(1)$:

(i) By Theorem 6.4, every seed node created before time step T_0' has a degree at most $O(\log n)$ contributed by global edges, and

(ii) All but $O(\log n)$ seed nodes created in time interval $[T_0', T_0]$ have a degree $O(1)$ contributed by global edges.

Note that the threshold of each node is chosen randomly and uniformly. Then, the communities that are created in these two time slots and satisfy the above conditions are vulnerable with probability $O(\log^{1-(a/2)} n)$ and $O(\log^{-b_1/2} n)$, respectively.

By Theorem 6.1(1), with probability $1 - o(1)$, there are at most $O\left(\frac{2n}{\log^a n}\right)$ communities. By the choice of $a > 6$, $-b_1/2 > 1 - (a/2)$ holds. Therefore, the expected number of vulnerable communities created before time step T_0 is $O\left(\frac{n}{\log^{a+(b_1/2)} n}\right)$.

Noting the independence of choice of threshold for each node, by using the Chernoff bound again, the lemma follows. □

By the proof of Lemma 10.4, there are only $O\left(\frac{n}{\log^{a+b_0} n}\right)$ communities born after T_0. So the total number of vulnerable communities in G is $O\left(\frac{n}{\log^{a+b_0} n} + \frac{n}{\log^{a+(b_1/2)} n}\right) = O\left(\frac{n}{\log^{a+b_0} n}\right)$.

Consider the infection priority tree T again. For any initial targeted set S of size $\lceil \log^c n \rceil$, the size of $\inf_H^{\mathrm{U}}(S)$ is at most

$$O\left(\left(\lceil \log^c n \rceil + \frac{2n}{\log^{a+b_0} n}\right) \cdot \log n \cdot \log^{a+1} n\right) = o(n).$$

This completes the proof of Theorem 6.10. □

11. EXPANDER CORE THEOREM

In this section, we establish Theorem 6.11.

Proof. (Proof of Theorem 6.11.) Suppose that $K = (V_K, E_K)$ is the induced subgraph of all the seed nodes of G. Then, V_K is the set of all the seed nodes of G, and E_K is the set of all the edges among the seed nodes of G. As in the proof of Theorem 6.1, E_K consists of two types of edges. The first-type edges are those created by Step 3(b) of Definition 5.1, and the second-type edges are those created by Step 3(c) of Definition 5.1. Clearly, the number of the first-type edges is small.

To prove our theorem, we will reduce the problem of bounding the conductance to the problem of bounding edge expansions, which is usually relatively easy.

Let $N = |V_K|$ be the size of K, and index by i ($1 \leq i \leq N$) the ith seed node generated in K. For a subset $S \subseteq V_K$, denote by $\Psi(S)$ the edge expansion of S, that is, $\Psi(S) = |\partial(S)|/|S|$, where $\partial(S)$ denotes the set of edges that have exactly one endpoint in S.

We rearrange the set S with the ordering as the time order of the creation of the nodes. There are two scenarios:

Case 1. Newly born node v is assigned to S.

In this case, the volume of S, denoted by vol(S), increases by at most $2d$, because there are at most d edges that are created between v and the existing nodes in K.

Case 2. Newly born node v is outside of S.

In this case, the increment of vol(S), the volume of S in K, is counted in $\partial(S)$.

Therefore, in any case, we have that

$$\mathrm{vol}(S) \leq 2d \cdot |S| + \partial(S).$$

According to this argument, the conductance of S in K satisfies

$$\Phi(S) = \frac{|\partial(S)|}{\mathrm{vol}(S)} \geq \frac{|\partial(S)|}{2d|S| + |\partial(S)|} = \frac{\Psi(S)}{2d + \Psi(S)}. \tag{11.1}$$

Recall that for an arbitrarily given graph $G = (V, E)$, the edge expansion $\Psi(G)$ and the conductance $\Phi(G)$ of graph G are defined as follows:

$$\Psi(G) = \min_{S:\ |S| \leq |V|/2} \Psi(S),$$

and

$$\Phi(G) = \min_{S:\ \mathrm{vol}(S) \leq \mathrm{vol}(G)/2} \Phi(S).$$

Consider a subset S of nodes in K whose volume is at most $\mathrm{vol}(K)/2$. If $|S| \leq N/2$, then $\Phi(S) \geq \Psi(S)/[2d + \Psi(S)]$. So, if $\Psi(S) \geq \alpha$, then $\Phi(S) \geq \frac{\alpha}{2d+\alpha}$. Otherwise, $|S| > N/2$, and so the complement of S, denoted by $\overline{S} = V_K \setminus S$, has a size less than $N/2$, but a volume at least $\mathrm{vol}(K)/2$. In this case, because $\partial(S) = \partial(\overline{S})$, we have $\Phi(S) \geq \Phi(\overline{S})$. Because (11.1) holds for every S regardless of its size, if $\Psi(\overline{S}) \geq \alpha$, then both $\Phi(\overline{S})$ and $\Phi(S)$ are at least $\frac{\alpha}{2d+\alpha}$. Thus, we need to show only that for any S of size at most $N/2$, its edge expansion $\Phi(S) \geq \alpha$.

To prove the edge expansion result of K, we further simplify the problem.

Let $K_2 = (V_K, E_2)$ be the graph obtained from K by deleting all the edges of the first type. Then, it is easy to see that

$$\Psi^K(S) \geq \Psi^{K_2}(S)$$

holds for every set $S \subset V_K$, where $\Psi^H(S)$ is the edge expansion of S in graph H.

The result follows easily from the construction of K and K_2.

Clearly, K_2 is a network constructed by the following uniform attachment model: whenever a seed node $v \in K$ is created, we create $d - 1$ edges from v to the existing nodes in K, each of which is randomly and uniformly selected.

Therefore, it suffices to use a constant lower bound for the edge expansion of the graph K_2. We will show that there is a constant β such that with probability $1 - o(1)$, for a graph H constructed from the uniform attachment model, $\Psi(H) \geq \beta$.

To prove our result, we recall a closely related result: [35] have shown that the networks of the preferential attachment model have a constant lower bound for the edge expansion of the networks. Our theorem shows the same lower bound for the networks of the uniform attachment model. We prove our theorem by following the approach in [35].

However, for our proofs, we consider a uniform recursive tree $T_{\overline{d}N}$ of size $\overline{d}N$ constructed as follows. We will use the idea of minivertices introduced in [35].

We define T_k to be the tree at the end of time step k in which the vertices of the tree are called minivertices. At time step 1, the tree is a single minivertex, and at time step k for $2 \leq k \leq \overline{d}N$, a minivertex with a single edge arrives and attaches randomly to a uniformly chosen minivertex in T_{k-1}.

Define a graph K' by first constructing the tree $T_{\overline{d}N}$ and then, for $1 \leq i \leq N$, contracting minivertices $\overline{d}i - j$ for $0 \leq j \leq \overline{d} - 1$, in which self-loops and multiedges are preserved. Comparing K' with K, the only difference is that self-loops in K' are permitted whereas there are not self-loops in K; K is equivalent to the case that, in the construction of K', when minivertex k ($k > \overline{d}$) arrives, the minivertex to which it attaches is uniformly chosen among those from 1 to $\overline{d} \cdot \lfloor k/\overline{d} \rfloor$ and the self-loops on the first node are deleted. It is easy to observe that the edge expansion of K is at least that of K' because self-loops can only make the edge expansion smaller. So, we have to show only that $\Phi(K') \geq \alpha$.

To this end, we will show that for any fixed integer $s \leq N/2$ and fixed subset $S \subseteq [N]$ of size s, the probability that $\Phi(S) < \alpha$ is small and the lemma follows from the union bound of all choices of S. For each subset $S \subseteq [N]$ and minivertex $k \in [\overline{d}]N$, we say that k is associated with S if and only if $k = \overline{d}i - j$ for some $1 \leq i \leq N$, $0 \leq j \leq \overline{d} - 1$ and $i \in S$. Fix a subset S of size $s \leq N/2$. We say that a minivertex is good if the edge created with its arrival belongs to $\partial(S)$. Otherwise, it is bad. That is, a minivertex is good if and only if its edge contributes to the cut $(S, \overline{S})$ and, thus, to the edge expansion of S. To prove that the probability that $\Phi(S) < \alpha$ is small, we will show that for any fixed subset $A \subseteq [\overline{d}N]$ of minivertices with size $|A| \leq \alpha s$, the probability is small that all minivertices associated with A are good and all those associated with $\overline{A} = [\overline{d}N \setminus A]$ are bad. Then, the union bound leads to the result. Formally, we prove the following lemma:

Lemma 11.1. *For any fixed $S \subseteq [N]$ and $A \subseteq [\overline{d}N]$ with $|S| = s$ and $|A| \leq \alpha s$, the probability that all minivertices in A are good and all those in $\overline{A}$ are bad is at most $\binom{\overline{d}s}{\alpha s} / \binom{\overline{d}N-\alpha s}{\overline{d}s-\alpha s}$.*

Proof. Let $A = A_1 \cup A_2$, where A_1 is the set of minivertices associated with S, and A_2 is the set of minivertices associated with $\overline{S}$. Let $|A_1| = k_1$ and $|A_2| = k_2$. Then we divide the minivertices in $\overline{A}$ into two parts according to their associations with S and $\overline{S}$. Note that $|\overline{A}| = \overline{d}N - k_1 - k_2$. Let $x_1 < x_2 < \cdots < x_{\overline{d}s-k_1}$ be the minivertices in $\overline{A}$ that associated with S and $x'_1 < x'_2 < \cdots < x'_{\overline{d}N-\overline{d}s-k_2}$ be the minivertices in $\overline{A}$ that associated with $\overline{S}$. For $1 \leq i \leq \overline{d}s - k_1$, let y_i be the number of minivertices in A that arrived prior to x_i and z_i be the number of minivertices in $\overline{A}$ that arrived prior to x_i. Symmetrically, for $1 \leq i \leq \overline{d}N - \overline{d}s - k_2$, let y'_i be the number of minivertices in A that arrived prior to x'_i and z'_i be the number of minivertices in $\overline{A}$ that arrived prior to x'_i. So, $x_i = y_i + z_i + 1$ and $x'_i = y'_i + z'_i + 1$. When x_i arrives, suppose that, at this stage, all minivertices in A are good and all those in $\overline{A}$ are bad. Note that the size of the tree at this stage is $x_i - 1$ and the number of minivertices associated with S is at most $i - 1 + y_i$ because $x_1, \ldots, x_{i-1}$ are all bad minivertices associated with S and at most y_i good minivertices are associated with S. Thus, because $y_i \leq |A|$, we have

$$\Pr[x_i \text{ is bad}] \leq \frac{i-1+y_i}{x_i-1} \leq \frac{i-1+y_i}{z_i+y_i} \leq \frac{i+|A|}{z_i+1+|A|}.$$

For the same reason, when x'_i arrives,

$$\Pr[x'_i \text{ is bad}] \leq \frac{i-1+y'_i}{x'_i-1} \leq \frac{i-1+y'_i}{z'_i+y'_i} \leq \frac{i+|A|}{z'_i+1+|A|}.$$

Recall that z_i (resp. z'_i) represents the number of minivertices in $\overline{A}$ that arrived prior to x_i (resp. x'_i) and x_i (resp. x'_i) is also in $\overline{A}$, the union of $\{z_i+1\}_{1\leq i\leq \overline{d}s-k_1}$ and $\{z'_i+1\}_{1\leq i\leq \overline{d}N-\overline{d}s-k_2}$ is exactly $\{1, 2, \ldots, |\overline{A}|\} = [\overline{d}N - |A|]$. We have that the probability, denoted by P, that all minivertices associated with A are good and all those associated with $\overline{A}$ are bad satisfies

$$\begin{aligned} P &\leq \prod_{i=1}^{\overline{d}s-k_1} \frac{i+|A|}{z_i+1+|A|} \cdot \prod_{i=1}^{\overline{d}N-\overline{d}s-k_2} \frac{i+|A|}{z'_i+1+|A|} \\ &= \frac{\prod_{i=1}^{\overline{d}s-k_1}(i+|A|) \cdot \prod_{i=1}^{\overline{d}N-\overline{d}s-k_2}(i+|A|)}{\prod_{i=1}^{\overline{d}N-|A|}(i+|A|)} \\ &= \frac{(\overline{d}s-k_1+|A|)! \cdot (\overline{d}N-\overline{d}s-k_2+|A|)!}{(\overline{d}N)! \cdot |A|!} \end{aligned}$$

$$
\begin{aligned}
&= \frac{(\bar{d}s + k_2)! \cdot (\bar{d}N - \bar{d}s + k_1)!}{(\bar{d}N)! \cdot |A|!} \\
&= \frac{(\bar{d}s)! \cdot (\bar{d}N - \bar{d}s)!}{(\bar{d}N - |A|)! \cdot |A|!} \cdot \prod_{i=1}^{k_1} \frac{\bar{d}N - \bar{d}s + i}{\bar{d}N - |A| + i} \cdot \prod_{i=1}^{k_2} \frac{\bar{d}s + i}{\bar{d}N - k_2 + i}.
\end{aligned}
$$

Because $|A| \le \alpha s \le \bar{d}s$, we have $\bar{d}N - \bar{d}s + i \le \bar{d}N - |A| + i$. Because $k_2 = |A_2|$, where A_2 is the set of minivertices in A associated with $\overline{S}$, and $\bar{d}N - \bar{d}s = \bar{d}|\overline{S}| \ge k_2$, we have $\bar{d}s + i \le \bar{d}N - k_2 + i$. Thus,

$$
\begin{aligned}
P &\le \frac{(\bar{d}s)! \cdot (\bar{d}N - \bar{d}s)!}{(\bar{d}N - |A|)! \cdot |A|!} \\
&= \frac{(\bar{d}s)!}{(\bar{d}s - |A|)! \cdot |A|!} \cdot \frac{(\bar{d}s - |A|)! \cdot (\bar{d}N - \bar{d}s)!}{(\bar{d}N - |A|)!} \\
&= \binom{\bar{d}s}{|A|} \Big/ \binom{\bar{d}N - |A|}{\bar{d}s - |A|}.
\end{aligned}
$$

Because $|A| \le \alpha s$ and $\bar{d} \ge 2$, we have $P \le \binom{\bar{d}s}{\alpha s} / \binom{\bar{d}N - \alpha s}{\bar{d}s - \alpha s}$. Lemma 11.1 follows. □

Lemma 11.1 means that for any fixed subset S of nodes of size s and any fixed subset A of edges of size at most αs in K', the probability (over the construction of K') that the cut $(S, \overline{S})$ is exactly A is at most $\binom{\bar{d}s}{\alpha s} / \binom{\bar{d}N - \alpha s}{\bar{d}s - \alpha s}$. Because there are $\binom{N}{s}$ choices for S and at most $\alpha s \cdot \binom{\bar{d}N}{\alpha s}$ choices for A when S is fixed, the probability that $\Phi(K') \le \alpha$ is given by the following union bound.

$$
\begin{aligned}
\Pr[\Phi(K') \le \alpha] &\le \sum_{s=1}^{N/2} \binom{N}{s} \cdot \alpha s \cdot \binom{\bar{d}N}{\alpha s} \cdot \frac{\binom{\bar{d}s}{\alpha s}}{\binom{\bar{d}N - \alpha s}{\bar{d}s - \alpha s}} \\
&\le \sum_{s=1}^{N/2} \alpha s \cdot \binom{\bar{d}N}{\alpha s} \cdot \frac{\binom{\bar{d}s}{\alpha s}}{\binom{(\bar{d}-1)N - \alpha s}{(\bar{d}-1)s - \alpha s}} \\
&\le \sum_{s=1}^{N/2} \alpha s \cdot \left(\frac{eN}{s}\right)^{\alpha s} \cdot \left(\frac{e\bar{d}}{\alpha}\right)^{\alpha s} \cdot \left(\frac{(\bar{d}-1)s - \alpha s}{(\bar{d}-1)N - \alpha s}\right)^{(\bar{d}-1)s - \alpha s} \\
&\le \sum_{s=1}^{N/2} \alpha s \cdot \left(\frac{N}{s}\right)^{\alpha s} \cdot \left(\frac{\bar{d}}{\alpha}\right)^{\alpha s} \cdot e^{2\alpha s} \cdot \left(\frac{s}{n}\right)^{(\bar{d}-1)s - \alpha s} \\
&= \sum_{s=1}^{N/2} \alpha s \cdot \left(\frac{e^2 \bar{d}}{\alpha}\right)^{\alpha s} \cdot \left(\frac{s}{n}\right)^{(\bar{d}-1-2\alpha)s}.
\end{aligned}
$$

Next, we have to show only that the function $f(s) = \alpha s \cdot \left(\frac{e^2\bar{d}}{\alpha}\right)^{\alpha s} \cdot \left(\frac{s}{n}\right)^{(\bar{d}-1-2\alpha)s}$ is of order $o(N^{-(c+1)})$ for any $2 \le s \le N/2$. Then, the above probability is $o(N^{-c})$ and Theorem 6.11 follows.

Let

$$f_1(s) = \alpha s,$$

$$f_2(s) = \left(\frac{e^2\overline{d}}{\alpha}\right)^{\alpha s}$$

and

$$f_3(s) = \left(\frac{s}{n}\right)^{(\overline{d}-1-2\alpha)s}.$$

So, $f(s) = f_1(s)f_2(s)f_3(s)$. Then the first derivative of $f(s)$ is

$$\begin{aligned} f'(s) &= f_1'(s)f_2(s)f_3(s) + f_1(s)f_2'(s)f_3(s) + f_1(s)f_2(s)f_3'(s) \\ &= \alpha f_2(s)f_3(s) \cdot \left[1 + \alpha s \ln\frac{e^2\overline{d}}{\alpha} + s(\overline{d} - 1 - 2\alpha)\left(1 + \ln\frac{s}{N}\right)\right]. \end{aligned}$$

Note that $f_2(s)$ and $f_3(s)$ are positive. Define function $g(s) = 1 + \alpha s \ln\frac{e^2\overline{d}}{\alpha} + s(\overline{d} - 1 - 2\alpha)\left(1 + \ln\frac{s}{N}\right)$. It is easy to calculate that for sufficiently large N, $g(2) < 0$ and $g(N/2) > 0$. So $f(s)$ decreases at $s = 2$ and increases at $s = N/2$. Because the second derivative of $g(s)$ is $g''(s) = (\overline{d} - 1 - 2\alpha)/s > (c + 1)/2 > 0$, $g'(s)$ is increasing and $g(s)$ is convex, which means that there is exactly one root for $g(s)$, denoted by $s = s_0$, in the interval $[2, N/2]$, and thus, $f(s)$ is monotonically decreasing in the interval $[2, s_0]$ and monotonically increasing in the interval $[s_0, N/2]$. So, the maximum value of $f(s)$ in the interval $[2, N/2]$ is determined by the maximum of $f(2)$ and $f(N/2)$, and we require that both of them are of order $o(N^{-(c+1)})$.

Since

$$f(2) = 2\alpha \cdot \left(\frac{e^2\overline{d}}{\alpha}\right)^{2\alpha} \cdot \left(\frac{2}{N}\right)^{2(\overline{d}-1-2\alpha)}.$$

$\alpha < \frac{\overline{d}-1}{2} - \frac{c+1}{4}$ implies $2(\overline{d} - 1 - 2\alpha) > c + 1$, and so $f(2) = o(N^{-(c+1)})$. Because

$$f(N/2) = \frac{\alpha N}{2} \cdot \left[\left(\frac{e^2\overline{d}}{\alpha}\right)^{\alpha} \cdot \left(\frac{1}{2}\right)^{\overline{d}-1-2\alpha}\right]^{N/2},$$

$f(N/2)$ decreases exponentially as N increases if $\left(\frac{e^2\overline{d}}{\alpha}\right)^{\alpha} < 2^{\overline{d}-1-2\alpha}$. Because the first derivative of function α^α is $\alpha^\alpha(1 + \ln\alpha)$ and it achieves minimum at $\alpha = 1/e$, the condition $\alpha < \frac{(\overline{d}-1)\ln 2 - 1/e}{2 + \ln(4\overline{d})}$ guarantees this. This completes the proof of Theorem 6.11. □

12. A PROTOCOL FOR SECURING COMPUTER NETWORK

Theorems 6.9 and 6.10 imply the following three discoveries:

1. Structures are essential to the security of networks against cascading failure models of attacks.

2. There is a trade-off between the role of structures and the role of thresholds in security of networks.

   Remark: There are ways to resist cascading failure of attacks. For example, we may define the threshold $\phi(v) = 1$ for all the nodes $v \in V$. However, this is expensive or even impossible in real-world applications, because large $\phi(v)$ means a high cost to protect node v. Our theory shows that if the network G is wellstructured, then small thresholds $\phi(v)$ for all the v's are sufficient to resist the network from cascading failures of a small-scaled attacks.
3. Neither power-law distribution [4] nor small-world property [48] is an obstacle of security of networks.

The first discovery is a mathematical principle. From the viewpoint of mathematics, we believe that structures determine the properties. In so doing, a structural theory of networks would provide provable guarantee for some of the key applications of network science. The nature of networks are the networks themselves, instead of just statistical measures of the networks. The investigation of interactions and structures of interactions of networks is, hence, essential to network theory and applications.

The second discovery explores that security of networks can be achieved theoretically by structures of networks, and that there is a trade-off between the role of structures and the role of lifting of the thresholds. This discovery is in sharp contrast to the current practice of network security engineering, which basically lifts the thresholds. Exploring the trade-offs between the role of structures and the role of thresholds in security of networks would provide a foundation for network security engineering, and hence, it would be exactly the subject of security theory of networks. Our discovery here plays such a role.

The third discovery is also highly nontrivial. The reasons are as follows intuitively speaking, power law allows us to attack a small number of top-degree nodes to generate a global cascading failure, and the small-world property means that spreading is so easy and so quick, that a small number of attacks might easily generate a global cascading failure.

Our discoveries imply that structure is a new, essential, and guaranteed source for security, and that the trade-off between the role of structure and the role of thresholds might provide both a full understanding of security and new technology for security engineering of networks.

We notice that the proofs of the security theorems, Theorems 6.9 and 6.10, are complicated systems of a number of topological, statistical, and combinatorial principles of networks of our security model. The system of the principles and proofs actually explore a criterion (or sufficient and necessary conditions) for security of networks. In Subsections 12.1 and 12.2, we will establish the criterion of network security, which is hidden in the proof system of our security theorems.

12.1. Sufficient Conditions for Security

The security model shows that dynamic and scale-free networks can be secure, for which homophyly, randomness, and PA are the underlying mechanisms, providing a principle for investigating the security of networks theoretically and generally.

Our security theorems explore some new discoveries between the roles of structures and of thresholds in the security of networks. The proofs of the security theorems provide a general framework to analyze both theoretically and practically security of networks.

It seems surprising that networks of the security model satisfy simultaneously all the properties stated in the four theorems, i.e., Theorems 6.1, 6.4, 6.6 and 6.8, and that a merging of the principles in Theorems 6.1, 6.4, 6.6, and 6.8 gives rise to the proofs of the security theorems, Theorems 6.9 and 6.10.

The proofs of Theorems 6.9 and 6.10 consist of a series of combinations of the the properties of networks generated by the natural selection of homophyly/kinship, randomness, and PA. By the proofs of Theorems 6.1, 6.4, 6.6, 6.8, 6.9, and 6.10, we can summarize some general principles for the community structure of networks that guarantees the security of the networks against cascading failure of any attacks of small scale.

Theorem 12.1. **(Structural principles of the security model.)** *Let $G = (V, E)$ be a network generated by our security model. Suppose that $\mathcal{N} = \{X_1, X_2, \cdots, X_N\}$ is the community structure of homochromatic sets of G, in which each X_i is called a natural community. Then, the natural community structure $\mathcal{N}$ of G satisfies the following properties:*

(1) (Overlapping free) The communities are disjoint.

(2) (Small-community phenomenon) The sizes of the communities are bounded by $\log^{O(1)} n$, where $O(1)$ is a constant, and n is the number of nodes of the network.

(3) (Inclusiveness and exclusiveness) There is no edge between nonseed nodes of distinct communities.

(4) (King node principle) With high probability, a community has a seed node that has degree significantly larger than any nodes of its own community and that represents the majority of external links of its own community.

(5) (Uniqueness) For a community X, the set of neighbors of X that are outside of X and are nonseed nodes of any community are all in a single community.

(6) (Robustness) Most nodes of a community have neighbors only within their own community.

(7) (Stability) The degree of a node is contributed mainly by nodes of its own community.

(8) (External decentrality) Both neighbors of a seed node outside of its own community and neighbors of nodes of a community outside of the community are homogenously distributed among different communities.

Recall that a community is strong if it will never be infected by the collection of outside communities, unless it has already been infected by nodes in the community itself.

(9) (Strong communities) Most communities are strong.

(10) (Infection priority tree) By modulo the small communities, we can extract an infection priority tree of the network, which is a directed tree in which all the directions are pointing to the roots.

(11) (Principle of infection among communities) Infections among the strong communities must be triggered by an edge in the infection priority tree of the network.

(12) (Directness) A seed node links to a nonseed node of a distinct community only if the seed node is created after the birth of the nonseed node.

(13) (Infection priority tree principle) The infection priority tree of the network has height $O(\log n)$.

Proof. By the proofs of Theorems 6.1, 6.4, 6.6, and 6.8.

□

A foundation for community analysis of the security of network is provided by (1)-(9), and (10) ensures that there is a huge number of strong communities. The existence of the infection priority tree T is the key to our proofs of the security theorems; (11), (12) and (13) ensure that cascading procedure among the strong communities has a path of length $O(\log n)$.

By the proofs of Theorems 6.9, and 6.10, we notice that the properties (1)–(13) in Theorem 12.1 are sufficient for the proofs of the security theorems 6.9, and 6.10.

We will show that each of the properties (1)–(13) in Theorem 12.1 is essential to the resistance of networks against cascading failures of networks. It is easy to see that (1) in Theorem 12.1 is a necessary condition. In Subsection 12.2, we will show that each of (2)-(13) in Theorem 12.1 is necessary for the security of networks.

12.2. Remarks: The Necessary Conditions for Security

In this subsection, we demonstrate that each of the properties (2)–(13) in Theorem 12.1 is necessary for security of networks.

Let $G = (V, E)$ be a network. Suppose that $\mathcal{N} = \{X_1, X_2, \cdots, X_N\}$ is a disjoint community structure of G. We prove by contradiction that each of Theorem 12.1(2)-(13) is necessary for security of networks. We consider the following cases.

Case 1. Theorem 12.1(2) is necessary for security of networks.

Suppose to the contrary that (2) fails to hold. We consider two subcases.

Subcase 1A. The X_i's are too small.

In this subcase, the structure $\mathcal{N}$ fails to represent the community structure of G.

Subcase 1B. The X_i's are too large.

In this subcase, if the X_i's are generated by a PA scheme as that in the security model, then even a single infected node in a community X_i might infect a large fraction of the community X_i. Because the X_i's are large, it is possible that a few infected communities could constitute a global failure of the network.

Therefore, if the communities are either too small or too large, then the community structure fails to resist cascading failure of the network.

Case 2. Theorem 12.1(3).

Suppose to the contrary that (3) fails to hold.

Given a community X, we call the nodes in X *vulnerable* if they are easily infected; these are usually the low-degree nodes in X.

For two disjoint communities X and Y, if there are edges between vulnerable nodes in X and vulnerable nodes in Y, then the failure of a single vulnerable node in X might easily infect a vulnerable node in Y and then the whole of Y.

By the assumption, there are many such edges between the vulnerable nodes of distinct communities. Therefore, the network fails to resist cascading failures by a small number of attacks. Therefore, the inclusiveness and conclusiveness in Theorem 12.1(3) are necessary for security of networks.

Case 3. Theorem 12.1(4).

The king node principle is necessary for security of networks.

By the arguments in Case 1 and Case 2, if a community fails to have strong nodes, then either the community is easily infected through external links, or the community is isolated from the network. Therefore, a community needs strong nodes (or king nodes) to lead the community and to communicate to nodes outside of the community.

Case 4. Theorem 12.1(5).

Suppose to the contrary that X, Y, and Z are three disjoint communities such that the strong nodes in X link to the vulnerable nodes in both Y and Z.

In this case, even if there is only one infected node in X, the following cascading failures could occur:

- All nodes in X are infected due to the heterogeneity of X.
- The vulnerable nodes in both Y and Z linking to nodes in X are all infected.
- All the nodes in Y and Z are infected.

Therefore, a single infected node in X might cause two infected communities Y and Z. This cascading procedure might continue from Y and Z, leading to an exponentially increasing of infected communities. This shows that Theorem 12.1(5) is necessary for security of networks.

This explains the reason why we create only one edge by the PA for a seed node in Definition 5.1.

Case 5. Theorem 12.1(6).

Let X be a community. By the arguments in Case 3, there are a few strong nodes in X, and most nodes in X are vulnerable nodes. By the arguments in Case 4, most external edges of a strong node link to strong nodes of other communities. Therefore, the most vulnerable nodes in X can link only to nodes in X itself.

Case 6. Theorem 12.1(7).

Let X be a community. By the arguments in Case 5, vulnerable nodes in X have this property, obviously.

Let x_0 be a strong node of X. Suppose to the contrary that the degree of x_0 is significantly contributed by nodes outside of X. In this case, x_0 is easily infected by its external neighbors. This makes X a vulnerable community.

Case 7. Theorem 12.1(8).

By the arguments in Cases 3, 4, and 5, every community has only a few strong nodes, the external links of a community are mainly through its strong nodes, and external edges of a strong node mainly link to strong nodes of other communities. Therefore, the external links of a community X are homogenously distributed among the strong nodes of other communities.

External links of a community are homogenously distributed in different communities.

By the arguments in Case 2, a strong node of a community has external links mainly to strong nodes of other communities. By definition, a strong node is unlikely to be infected by the collection of its external neighbors, meaning that the external neighbors of a strong node cannot be too many and must be homogeneously distributed.

Case 8. Theorem 12.1(9) is necessary.

Suppose to the contrary that many communities' X_i's are vulnerable. Then clearly the community structure $\mathcal{N}$ fails to resist cascading failure of the network.

Case 9. Theorem 12.1(10), (11), and (12).

By the arguments in Case 4, there exists an infection priority tree T such that an infection from a strong community X to a distinct strong community Y occurs only through an edge of tree T. Suppose to the contrary that the infection priority tree T is undirected. In this case, there are arbitrarily long infecting paths in T.

Therefore Theorem 12.1(10), (11), and (12) are all necessary for security of networks.

Case 10. Theorem 12.1(13).

Suppose to the contrary that the infection priority tree T has a long path from a leaf to the root. In this case, a single infected node might cause a long chain of infected communities. Therefore, Theorem 12.1(13) is necessary for security of networks.

We thus conclude that each of the properties in Theorem 12.1 is necessary for security of networks.

By the arguments in Subsections 12.1 and 12.2, we have that a community structure $\mathcal{N}$ of a network G resists cascading failure of G by attacks if and only if all the properties in (1)–(13) of Theorem 12.1 are satisfied.

This explains the reason why our model in Definition 5.1 proceeds in a carefully organized way as it is. In particular, in the definition, we require that a seed node can use the PA to create ONLY one edge. (We simply cannot create more than 1 such edge, in engineering, there is no such edge in fact.) This again indicates that security is an engineering requirement, which can only be achieved carefully, instead of arbitrarily or freely.

Therefore, community structure does resist cascading failure of attacks if and only if a number of combinatorial principles are satisfied. The result demonstrates that the role of community structure in security of networks is conditional, contrary to the naïve intuition that a community may protect members of its community [42, 50], and that there is no simple statistical measure of networks that guarantee the security of networks.

12.3. Computer Network Engineering

Currently, most models of networks have a background of social interactions and nature evolving. Consequently, the networks generated by the models are usually far from the real engineering networks, such as the internet topology. In a recent survey, [49] proposed a network modeling of reverse-engineering. This explored a community structure of the real Internet with all the properties as stated in our principles (1)–(13) in Theorem 12.1, except that the nodes responsible for global links have small degrees. This property is easy to achieve by a reduction from the networks of our model. It proceeds as follows:

Let $G = (V, E)$ be a network generated by our model. Let X be a natural community of G and x_0 be the seed node of X. We replace x_0 by an edge $(x_0^{\text{in}}, x_0^{\text{out}})$ such that x_0^{in} and x_0^{out} play the internal and external roles of x_0, respectively. That is, every edge from a node $x \in X$ to x_0 is replaced by an edge (x, x_0^{in}), and every edge from a node $y \notin X$ to x_0 is replaced by an edge from y to x_0^{out}. Let H be the network obtained by extending the seed nodes of G by the reduction above. Then H is very similar to a network by the reverse-engineering introduced in [49].

By construction, H consists of some small communities, each of which has a base station that is central to its own community, and a representative node that links to the central node and to nodes outside of the community. Now, the global core of the network is basically the part of external links of the seed nodes, which are small, and each of which has a small degree.

However, in this case, H is no longer secure. We demonstrate that the security proofs of G cannot be carried to H. The reasons are as follows. In H, for every community X and its seed node x_0, we replace it by a link $(x_0^{\text{in}}, x_0^{\text{out}})$. In this case, we still regard both x_0^{in} and x_0^{out} as being in the community X. We say that a node y is strong if the degree of y contributed by its neighbors outside its community is negligible compared to the degree of y. From this understanding, x_0^{in} is still a strong node, but x_0^{out} is no longer a strong node. In fact, we may regard x_0^{out} as the *boundary of community* X. Then, almost all external links of a community are through the boundary of the community. Now, the following infection of nodes among different communities occurs with high probability. Suppose that X is a community and its boundary x_0^{out} has been infected. By the construction of H, it is highly likely to have a nonseed node y in a community $Y \neq X$ such that there is an edge between x_0^{out} and y. Because x_0^{out} is infected, and y is nonseed, it is highly likely that y is infected. Furthermore, many nodes in Y, including the pair y_0^{in} and y_0^{out} generated from the seed y_0 in G, are infected. Because all the boundary nodes are weak, y_0^{out} could easily infect all its neighbor boundary nodes such as $a_0^{\mathrm{out}}, b_0^{\mathrm{out}}, \cdots$. Furthermore, all the infected boundary nodes $a_0^{\mathrm{out}}, b_0^{\mathrm{out}}, \cdots$ simultaneously infect its vulnerable nonseed nodes outside their own communities. This shows that H is no longer secure.

12.4. A Protocol for the Security of Computer Network

By the arguments in Subsection 12.3, the reduced network H from the network G of the security model is insecure, so cannot be the model of a real-world computer network. However, there is an easy way to secure H. Let H' be the graph obtained from H by deleting all the edges between boundary nodes and nonseed nodes outside the communities of the boundary nodes. For H', the obstacle of security for H disappears, and the proofs of security of G can be carried out to H'. This gives rise to a model of a computer network that is guaranteed to be secure.

Therefore, our principles (1)–(13) in Theorem 12.1 reflect the phenomena of engineering networking.

The reduction above also suggests a generator of networks that very well simulates the engineering of computer networks and guarantees the security of the networks.

12.5. Expander core theorem of computer network H'

An important property of H' is the following result.

Theorem 12.2. **(Expanding core theorem.)** *For $a > 0$, and $d \geq 3$, let G be a network of the security model with n nodes, affinity exponent a and average number of edges d. Let H and H' be the networks constructed from G in Subsections 12.3 and 12.4. Let K be the induced subgraph of all the boundary nodes x^{out} of H'. Then, K is an expander, or equivalently, there is a large constant $\alpha > 0$ such that for every subset S of the nodes of K, the conductance of S in K is greater than or equal to α.*

Proof. By Theorem 6.11. □

By Theorem 12.2, communications in K are easy and quick.

Therefore, our theory provides a foundation for the principles of computer network security. However, practice principles based on our theory must be built based on computer network experiments, for which one of the main tasks is to determine the appropriate affinity exponent a for the networks with real-world sizes. This problem is not solved by our theory, because our security theorems assume the affinity exponent $a > 4$ or even > 6, and even in this case, we assume that $\log^a n$ is negligible with n. This assumption fails to apply to the real-world computer networks.

13. CONCLUSIONS AND FUTURE DIRECTIONS

We proposed the definition of security of networks, and the security model of networks. It was shown that with appropriate choice of affinity exponent, the networks generated by our model are provably secure. We also proposed a protocol for securing a computer network based on our model of networks. It is conceivable that our theory, a dynamic theory of networks, provides a foundation for a security theory of networks in a wide range of areas such as computer engineering, communication engineering, industry, nature, society, economics, finance, and Internet finance, etc., to name a few. Our theory suggests a number of new directions for future network study, for which we introduce some of the important issues as follows:

(1) Security protocols of industry networks, economics networks, finance, and Internet finance networks, and the security criterion of computer networks.

For security engineering of networks, the most important issue would be to define a measure to determine the Network Security Index (NSI). This criterion, once defined, allows us to quantitatively understand the security performance of a given network and to establish a combinatorial optimization theory for security of networks.

(2) The roles of the affinity exponent a in the dynamics of real-world networks, including the theory of resisting cascading failures, the theory of emergence of cooperation in evolutionary games, and the principles of self-organization and social organization in networks.

In the real world, the networks are finite graphs, although large. For the networks of the security model with sizes bounded by a large number N, the dynamics of the networks are determined by the affinity exponent a. In this case, if a is too small or too large, the networks of the model might fail to satisfy the dynamics desired. The issue is to characterize the desired affinity exponent a with fixed number N of nodes of the networks and to develop new applications based on the analyses.

(3) Evolutionary game theory on the basis of the security model

One of the most important topics of network dynamics is the theory of evolutionary games, because the security model follows some of the ideas of Darwin's natural selection [12]. The model may provide a foundation for an evolutionary game theory.

(4) Robustness of networks

Robustness is closely related to, but certainly different from, our security theory, here. It would be very interesting to develop a robustness theory of networks.

ACKNOWLEDGMENTS

The authors are grateful to the referee for reviewing this long study and for helpful suggestions for revising the article.

The first author thanks his former and current students Wei Zhang and Xiaohui Zhang for help with the experiment in Figures 1–4 in the article.

FUNDING

The authors are partially supported by the Grand Project "Network Algorithms and Digital Information" of the Institute of Software, Chinese Academy of Sciences, by an

NSFC grant No. 61161130530, and by a China Basic Science Program (973) Grant No. 2014CB340302.

REFERENCES

[1] R. Albert, H. Jeong, and A. L. Barabási. “Error and Attack Tolerance of Complex Networks” *Nature* 406 (2000), 378–381.

[2] R. M. Andersen and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford University Press, 1991.

[3] A. L. Barabási. “Scale-Free Networks: A Decade and Beyond.” *Science* 325 (2009), 412–413.

[4] A. L. Barabási and R. Albert. “Emergence of Scaling in Random Networks.” *Science* 286 (1999), 509–512.

[5] A. Blum, T. Chan, and M. R. Rwebangira. “A Random-Surfer Web Graph Model.” In *Proceedings of the Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pp. 238–246. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2006.

[6] L. Blume, D. Easley, J. Kleinberg, R. Kleinberg, and E. Tados. “Which Networks are Least Susceptible to Cascading Failures?” Paper presented at the 52nd Annual Symposium on Foundations of Computer Science, Palm Springs, CA, October 22–25, 2011.

[7] B. Bollobás. *Random Graphs*. Cambridge, UK: Cambridge University Press, 2001.

[8] Bollobás, B. and O. Riordan. “The Diameter of a Scale-Free Random Graph.” *Combinatorica* 24 (2004), 51C34.

[9] H. Chernoff. “A Note on an Inequality Involving the Normal Distribution.” *The Annals of Probability* 9 (1981), 533–535.

[10] F. Chung and L. Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

[11] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. “Resilience of the Internet to Random Breakdowns.” *Physical Review Letters* 85:21 (2000), 4626–4628.

[12] C. Darwin. *On the Origin of Species by Means of Natural Selection*. London, UK: John Murray, 1859.

[13] S. Dommers, R. van der Hofstad, and G. Hooghiemstra. “Diameters in Preferential Attachment Models.” *J Stat Phys* 139 (2010), 721C107.

[14] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Oxford, UK: Oxford University Press, 2010.

[15] P. Erdös and A. Rényi. “On Random Graphs.” *Publ. Math.* 6 (1959), 290–297.

[16] P. Erdös and A. Rényi. “On the Evolution of Random Graphs,” *I. “Magyar Tud. Akad. Mat. Kutató Int. Kózl.* 5 (1960), 17–61.

[17] J. Goldenberg, B. Libai, and E. Muller. “Talk of the Network: A Complex Systems Look at the Underlying Process of the Word-of-Mouth.” *Marketing Letters* 12:3 (2001), 211–223.

[18] J. Goldenberg, B. Libai, and E. Muller. “Using Complex Systems Analysis to Advanced Marketing Theory Development.” *Academy of Marketing Science Review* 9:3 (2001), 1–18.

[19] M. Granovetter. “Threshold Models of Collective Behavior.” *The American Journal of Sociology* 83:6 (1978), 1420–1443.

[20] D. Kempe, J. Kleinberg, and E. Tardos. “Maximizing the Spread of Influence Through a Social Network.” SIGKDD, 137–146, 2003.

[21] J. Kleinberg. “Navigation in a Small World.” *Nature* 406 (2000), 845.

[22] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. “Stochastic Models for the Web Graph.” Paper presented at FOCS, Redondo Beach, CA, November 12–14, 2000.

[23] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. “Trawling the Web for Emerging Cyber-Communities.” *Proceedings of the 8th World Wide Web Conference*, pp. 403–416, Philadelphia, PA: Elsevier, 1999.

[24] J. Leskove and C. Faloutsos. "Scalable Modeling of Real Graphs Using Kronecker Multiplication." Paper presented at ICML, Corvallis, OR, June 20–24, 2007.

[25] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanation." Paper presented at ACM KDD, Chicago, IL, August 21–24, 2005.

[26] A. Li, J. Li, and Y. Pan. "Homophyly/Kinship Hypothesis: Natural Communities, and Predicting in Networks." *Physica A* 420 (2015), 148–163.

[27] A. Li, J. Li, and Y. Pan. "Discovering Natural Communities in Networks." *Physica A* 436 (2015) 878–896.

[28] A. Li, J. Li, Y. Pan, X. Yin, and X. Yong. Homophyly/Kinship Model: Naturally Evolving Networks. *Scientific Reports* to appear.

[29] A. Li, X. Li, Y. Pan, and W. Zhang. "Strategies for Security of Networks." *Science China, Informatics* 58 (2015), 012107:1–012107:14.

[30] A. Li and P. Peng. "Community Structures in Classical Network Models." *Internet Math.* 7:2 (2011), 81–106.

[31] A. Li and P. Peng. "Small Community Phenomenon in Networks." *Mathematical Structures in Computer Science* 22 (2012), 1–35.

[32] A. Li, X. Yin, and Y. Pan. "Three-Dimensional Gene Map of Cancer Cell Types: Structural Entropy Minimization Principle for Defining Tumour Subtypes." 2015, To appear.

[33] P. Mahadevan et al. "Systemic Topology Analysis and Generation Using Degree Correlations." Paper presented at Sigcomm, Pisa, Italy, September 11–15, 2006.

[34] H. Mahmoud and R. Smythe. "A Survey of Recursive Trees." *Theory of Probability and Mathematical Statistics* 51 (1995), 1–27.

[35] M. Mihail, C. H. Papadimitriou, and A. Saberi. "On Certain Connectivity Properties of the Internet Topology." *Journal of Computer and System Sciences*, 72 (2006), 239–251.

[36] S. Morris. "Contagion." *Review of Economic Studies* 67 (2000), 57–78.

[37] A. E. Motter. "Cascade Control and Defense in Complex Networks." *Physical Review Letters* 93:9 (2004), 098701.

[38] M. E. J. Newman. "The Structure and Function of Complex Networks." *SIAM Rev.* 45:2 (2003), 167–256 (electronic).

[39] B. Pittel. "Note on the Heights of Random Recursive Trees and Random m-ary Search Trees." *Random Structures and Algorithms* 5 (1994), 337–347.

[40] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. "Dynamical and Correlation Properties of the Internet." *Physical Review Letters* 87:25 (2001), 258701.

[41] E. Ravasz and A. L. Barabási. "Hierarchical Organization in Complex Networks." *Physical Review E* 67 (2003), 056104.

[42] M. Salathé and J. H. Jones. "Dynamics and Control of Diseases in Networks with Community Structure." *PLOS Comput Biol* 6:4(2010), e1000736.

[43] F. Schweitzer et al. "Economic Networks: The New Challenges." *Science* 325:24 (2009), 422–425.

[44] S. Tzu. *The Art of War*. Trans. R. D. Sawyer. Boulder, CO: Westview Press, 1994.

[45] R. Toivonen. "A Model for Social Networks." *Physica A: Statistical and Theoretical Physics* 371:2 (2006), 851–860.

[46] A. Vazquez. "Growing Network with Local Rules: Preferential Attachment, Clustering Hierarchy and Degree Correlation." *Physical Review E* 67 (2003), 026112.

[47] D. J. Watts. "A Simple Model of Global Cascades on Random Networks." *Proc. National Academy of Sciences, USA* 99:9 (2002), 5766–5771.

[48] D. J. Watts and S. H. Strogatz. "Collective Dynamics of Small World Networks." *Nature* 393 (1998), 440–442.

[49] W. Willinger and M. Roughan. "Internet Topology Research Redux." In *Recent Advances in Networking*, edited by H. Bonaventure, pp. 1–59, New York, NY: ACM, 2013.

[50] X. Wu and Z. Liu. "How Community Structure Influences Epidemic Spread in Social Networks." *Physica A: Statistical Mechanics and Its Applications.* 382:2 (2008), 623–630.