Deployment Aware Modeling of Node Compromise Spread in Wireless Sensor Networks Using Epidemic Theory

PRADIP DE, YONGHE LIU, and SAJAL K. DAS
University of Texas, Arlington

General Terms: Sensor Networks, Epidemic Theory

Additional Key Words and Phrases: Random Key Predistribution, Random Graph, Group-based Deployment.

1. INTRODUCTION

As wireless sensor networks are unfolding their vast potential in a plethora of application environments [Chong and Kumar 2003; Akyildiz et al. 2002], security still remains one of the most critical challenges yet to be fully addressed. In particular, a vital problem in the highly distributed and resource constrained environment is node compromise, where a sensor node can be completely captured and manipulated by the adversary. While extensive work has focused on designing schemes that can either defend and delay node capture or timely identify and revoke compromised nodes themselves [Chan et al. 2005], little attention has been paid to the node compromise process itself. Inspired by recently emerged viruses that can spread over air interfaces, and the various broadcast protocols for transferring data/executable code across the network, we identify in this paper the threat of epidemic spreading of node compromises in large scale wireless sensor networks. In essence, we present a model that captures the unique topological characteristics of typically deployed sensor networks in conjunction with pairwise key schemes, and identify the key factors determining the potential epidemic outbreaks that in turn can be employed to devise corresponding defense strategies.

1.1 Motivation

Due to its scarce resources and hence low defense capabilities, node compromises can be expected to be common phenomena for wireless sensor networks in unattended and hostile environments. While extensive research efforts, including those from ourselves [Chadha et al. 2005], have been engineered toward designing resilient network security

Authors' Address: Center for Research in Wireless Mobility and Networking (CReWMaN)
Department of Computer Science and Engineering,
The University of Texas at Arlington
416 Yates Street, Arlington, TX 76019.
Email: [pradip.de, yonghe, das]@uta.edu

mechanisms [Du et al. 2004; Chan et al. 2003; Liu and Ning 2003], the compromise itself and, in particular, the propagation of node compromise (possible epidemics) have attracted little attention.

While node compromise, thanks to physical capture and succeeding analysis, is naturally constrained by the adversary's capability, software originated compromises can be much more damaging. An example was the recently surfaced virus *Cabir*[1] that spread over the bluetooth air interface and created havoc among bluetooth enabled mobile phones. This instance clearly heralds the advent of viruses that can spread over-the-air. Wireless sensor networks, with their inherent properties of high densities and large scale deployments, coupled with the fact that they are generally deployed in mostly unattended terrains, are undoubtedly vulnerable to possible virus/malware outbreaks. Inescapably, viruses targeting wireless sensor networks will emerge and, consequently, node compromise by way of virus spreading (over the air interface) can potentially devastate the entire network in a short period of time.

In order to further motivate our work, below we examine a few special characteristics of a sensor network as opposed to other networks that underline its vulnerability.

—*resource constraints* : A sensor node is very limited in resources and indeed it requires little effort to render it inoperative. For instance, with its very little memory capacity, a sensor node is vulnerable to a malware that can generate data to consume all its memory space. Moreover, bombarding a sensor node with a deluge of packets can not only cause its buffers to overflow but can also cause its battery to drain out quickly.

—*large scale and high density* : Sensors are typically assumed to be deployed in significant numbers and in large scale (battlefields, agricultural lands, etc). In addition, the density of the network is also expected to be high. In such a scenario, a malware that can use the communication channel to spread, can propagate very fast and compromise the whole network causing devastating effects.

—*deployment areas* : A sensor network is envisioned as a network of embedded devices that would be deployed in terrains generally inaccessible. Thus, monitoring every part of the network often is impossible. It is only at the base station that data arriving from the sensor network can by analyzed. Furthermore, nodes are often not individually addressable and thus it can be impossible to troubleshoot individual devices. Thus, the difference between the time when a node is compromised and when detection takes place can be very high and in some cases, the damage would already have taken devastating proportions.

Recently, several protocols for reprogramming the sensors over-the-air have been proposed in the recent literature [Levis et al. 2004; Hui and Culler 2004; Wang 2004]. These protocols are extremely useful in re-tasking or reconfiguring the sensor network as a whole and deploying new applications on the devices over the air. However, these broadcast or code dissemination protocols can easily serve as vehicles in transferring a piece of malware across the whole network very quickly. Thus, malwares which do not have the ability to establish communication between neighboring nodes can exploit these protocols in transferring themselves to the whole network. The density and large scale nature of wireless sensor networks would only make matters worse and facilitate the fast propagation over these protocols

[1] http://www.f-secure.com/v-descs/cabir.shtml

Indeed, malware spreading over the Internet has been widely studied, and notably by means of epidemic theory [Staniford et al. 2002]. However, the marked difference in the topological aspects of the Internet with that of a sensor network often render those models unusable in the latter. The Internet is typically characterized as a scale free network showing both the properties of *preferential attachment* and *growth* which are largely missing in wireless sensor networks. Moreover, the distance and pairwise key based schemes used for securing the communication among sensor nodes further underlines the requirement for an epidemic model specific to sensor networks. Although there are several algorithms and protocols for data dissemination and routing in sensor network [Braginsky and Estrin 2002] that are based on epidemic principles, a consolidated formal model to quantify the propagation rate and other important parameters is yet to be designed.

1.2 Our Contribution

In this paper, we investigate the spreading process of node compromise in large scale wireless sensor networks. Starting from a single point of failure, we assume that the adversary can effectively compromise neighboring nodes through wireless communication and thus can threat the whole network without engaging in full scale physical attacks. In particular, due to security schemes employed by the sensor networks, we assume that communication can only be performed when neighboring nodes can establish mutual trust by authenticating a common key. Therefore, node compromise is not only determined by the deployment of sensor nodes which in turn affects node density, but also determined by the pairwise key scheme employed therein. By incorporating these factors of the networks, we propose an epidemiological model to investigate the probability of a breakout (compromise of the whole network) and if not, the sizes of the affected components (compromised clusters of nodes). Furthermore, we analyze the effect of node recovery in an active infection scenario and obtain critical values for these parameters that result in an outbreak. We focus our analysis on two specific types of node deployment scenarios, namely uniform random deployment and group based deployment of nodes (where the actual resident points of the nodes of a group are assumed to follow a particular spatial distribution about the group deployment point). Through extensive simulations, we not only show that our analytical results can closely capture the effects in a wide range of network setups, but also provide deeper insights on the temporal dynamics of the epidemic process under each deployment scenario.

The remainder of the paper is organized as follows. In Section 2 we present the preliminaries, including the threat model, random key pre-distribution, and epidemic theory. In Section 3, we study the compromise propagation without node recovery and with node recovery, and detail our analytical results. In section 4, we discuss the basic reproductive number, which is a very important parameter of Epidemic Theory, and how it is evaluated based on the network parameters. We perform experimental study in Section 5. Related work is presented in Section 6 and we conclude in Section 7.

2. PRELIMINARIES

In this section, we briefly provide a preliminary overview of a a set of topics relevant to the current work, namely pairwise key pre-distribution and epidemic theory. Subsequently, we delineate the threat model which provides the basis of the epidemic model for compromise spread.

2.1 Pairwise Key Pre-distribution

In this subsection, we briefly overview the pairwise key scheme for securing the communication between neighboring nodes in a sensor network based on key pre-distribution. This shared secret key based secure communication is especially popular in sensor networks owing to the prohibitive resource consumption of most public key cryptographic techniques.

Due to the severe resource constraint of wireless sensor networks and limited networking bandwidth, proposed pairwise key schemes have commonly adopted the pre-distribution approach instead of online key management schemes. The concept of pre-distribution was originated from [Eschenauer and Gligor 2002], where the authors propose to assign a number of keys, termed *key ring* randomly drawn from a key pool. If two neighboring nodes share a common key on their key rings, a shared pairwise key exists and a secure communication can be established. An enhanced scheme termed *Q-composite* was proposed in [Chan et al. 2003], where two neighboring nodes can establish secure communication only if at least Q keys are shared on their key rings. Pre-distribution schemes that rely on bivariate polynomials is discussed in [Liu and Ning 2003]. In this scheme, each sensor node is pre-distributed a set of polynomials. Two sensor nodes with the same polynomials can respectively derive the same key. A pairwise key predistribution scheme was proposed by the authors in [Du et al. 2005].

Regardless of the specific key distribution scheme, a common parameter capturing the performance is the probability that two neighbors can directly establish a secure communication or, in other words, share at least one key. We denote this key sharing probability by q. Thus, two physical neighbors can communicate securely with probability q. The factors on which q depends, such as the key pool size or the individual key ring sizes, have been studied in previous works [Eschenauer and Gligor 2002; Chan et al. 2003]. The value of q is crucial in controlling the degree of connectivity of the securely communicating sensor network. As we will reveal later, q plays an important role in the spreading of node compromise, as direct communication (as explained subsequently, in the threat model) can result in propagation of malicious code. A high value of q would make the network highly connected while at the same time increase the network's susceptibility to compromise propagation.

2.2 Epidemic Theory

Originally, epidemic theory concerns about contagious diseases spreading in the human society. The key feature of epidemiology [Anderson and May 1992; Pastor-Satorras and Vespignani 2001b; 2001a; May and Lloyd 2001; Hethcote 2000] is the measurement of infection outcomes in relation to a *population at risk*. The population at risk basically comprises of the set of people who possess a susceptibility factor with respect to the infection. This factor is dependent on several parameters including exposure, spreading rate, previous frequency of occurrence etc., which define the potential of the disease causing the infection. Various models have been proposed and thoroughly investigated in Epidemic theory that characterize the infection spreading process. Example models include Susceptible Infected Susceptible (SIS) Model, Susceptible Infected Recovered (SIR) Model etc. In the former, a susceptible individual acquires infection and then after an infectious period, (i.e., the time the infection persists), the individual becomes susceptible again. On the other hand, in the latter, the individual recovers and becomes immune to further infections.

Of particular interest is the non-equilibrium phase transition of the spreading process that is dependent on an epidemic threshold: if the epidemic parameter is above the threshold, the infection will spread out and become persistent; on the contrary, if the parameter is below the threshold, the virus will die out.

Epidemic theory indeed has been borrowed to the networking field to investigate virus spreading. However, a commonly adopted model is perfect mixing, where it is assumed that an infected node has equal probability of infecting any other node on the network. In wireless sensor networks, due to distance and security constrained communication pattern, such an assumption immediately becomes unrealistic. In this paper, we will mainly rely on random graph models to characterize the unique connectivity of the sensor network under different conditions of node deployment and perform the epidemic study [Stauffer 1985; Callaway et al. 2000].

2.3 Node Recovery

In the event that a node is compromised, its secrets will be revealed to the attacker. The network may attempt to *recover* the particular node. Recovery might be realized in several possible ways. For example, the keys of the nodes might be revoked and the node may be given a fresh set of secret keys. In this context, key revocation, which refers to the task of securely removing keys that are known to be compromised, has been investigated as part of the key management schemes, for example in [Chan et al. 2005]. Moreover, recovery can also be achieved by simply removing the compromised node from the network, for example by announcing a blacklist, or simply reload the node's programs. More sophisticated methods may include immunizing a node with an appropriate antivirus patch that might render the node immune from the same virus attack.

Regardless, in our analysis, we will study virus spreading under the two cases respectively depending on whether a node can be recovered or not.

2.4 Threat model

We assume that a single node or a very small set of nodes are initially compromised by an adversary. This compromise can be through physical capture of the device and subsequent analysis of the node resulting in installation of a malware and acquisition of its key ring. A compromised node can establish direct secure communication with any neighbor with which it shares at least one key. We assume that a compromised node, by directly communicating with a susceptible node through their shared secure communication channel, can transfer any malware, thereby spreading the infection and conducing to the compromise of the susceptible node. This process can repeat itself and ultimately lead to the compromise of the whole network or a significant portion of it.

As stated above, communication among sensor nodes is not only constrained by their distances, but also shall be secured and thus determined by the probability of pairwise key sharing. Therefore, the spreading of node compromise is dependent on the network deployment strategy and the pairwise key sharing scheme employed therein, based on which the network would be compromised to varying degrees.

Although a sensor network is significantly different from the Internet in its topological properties, the spread of node compromise in a sensor network, thanks to its dense nature, can also lead to an epidemic effect analogous to virus spreading over the Internet.

We consider this epidemic effect as the key threat to the network and hence the investigation target of this paper. To emphasize the severity of this threat, as an example, the

recent spread of the Sasser worm paralyzed 1.5 million computers all over the Internet. When coupled with the emergence of viruses that spread over the air, these phenomena point to the vulnerability staring at large and dense sensor networks consisting of resource limited nodes.

3. MODELING AND ANALYSIS OF COMPROMISE PROPAGATION

In this section, we analyze the propagation of node compromise originating from a single node that has been affected. Our focus is to study the outbreak point of the epidemic effect where the whole network will fall victim to the compromise procedure.

Our key method is to characterize the sensor network, including its key distribution, by mathematically formulating it as a random graph whose key parameters are precisely determined by those of the sensor network. Therefore, the investigation of epidemic phenomena can be performed on the random graph instead.

We perform our analysis on two types of sensor network topology models. In our first model, we assume that the sensor nodes are uniformly randomly distributed. In our second model, we assume a more realistic scenario where deployment knowledge is incorporated in the analysis. We assume that nodes are deployed in groups and the resident points of each node in a group follows a two dimensional gaussian distribution about the deployment point. Subsequently, given these two deployment approaches, we observe the epidemic process under two scenarios: without node recovery and with node recovery, depending on whether infected nodes will be recovered by external measures like key revocation, immunization, and so on.

Our goal is to analyze and compare epidemic spread progress and effect under different topological scenarios of the sensor network. In the following two subsections, we derive the degree distribution for the two random graph deployment models, viz. uniform random distribution and groups of gaussian random distribution.

3.1 Network Model

In this section, we will model the network topology of the overlay key sharing graph above the physical network. The outcome of our model is the degree distribution of the key sharing overlay topology. In our analytical derivation, we incorporate the deployment knowledge of the sensor nodes by using different deployment models to finally derive the network degree distribution.

3.1.1 *Uniform Random Distribution.* Assume that sensor nodes are uniformly randomly deployed in a region with area A. Let $\rho = \frac{N}{A}$ denote the node density of the network where N is the total number of the nodes. For a sensor node with communication range R, the probability that l nodes are within its communication range is given by

$$p(l) = \binom{N-1}{l} p^l (1-p)^{N-1-l} \tag{1}$$

where p is defined by

$$p = \frac{\pi R^2}{A} = \frac{\pi R^2 \rho}{N}. \tag{2}$$

Thus p is the probability of a link existing at the physical level, i.e., whether the two nodes fall within their respective communication ranges.

We further assume that the probability that two neighboring nodes sharing at least one key in the random pre-distribution pairwise key is q. Notice that q is determined by the specific pairwise key scheme employed. For a particular node having l neighboring nodes, the probability that there are k nodes, $k \leq l$, sharing at least one key with it is given by

$$p(k|l) = \binom{l}{k} q^k (1-q)^{l-k} \tag{3}$$

Therefore, with uniform random deployment, the probability of having k neighboring nodes sharing at least one key is

$$p_u(k) = \sum_{l=k}^{N-1} p(l)p(k|l) \tag{4}$$

$$= \sum_{l=k}^{N-1} \binom{N-1}{l} p^l (1-p)^{N-1-l} \binom{l}{k} q^k (1-q)^{l-k} \tag{5}$$

3.1.2 *Group based Deployment Model : Two Dimensional Gaussian Random Distribution.* The other deployment model that we consider in this paper is group based deployment [Du et al. 2004; Yu and Guan 2005]. In this model, sensors are divided into groups where each group is deployed (e.g., dropped from an airplane) at a particular location. Due to the uncertainty of the deployment procedure, sensor nodes within each group are often randomly distributed around the targeted deployment point. Specifically, we make the following assumptions about this model.

—N sensor nodes to be deployed are divided into t equal size groups each with n nodes. Each group, G_i, for $x_i = 1, \ldots, t$ and $y_i = 1, \ldots, n$, is deployed from the deployment point (x_i, y_i).

—The deployment points are assumed to be arranged in a grid, which is commonly assumed.

—During deployment, the resident points of the node k in group G_i with deployment point (x_i, y_i) follow probability distribution $f_k^i(x, y|k \in G_i)$. An example of this distribution is a two-dimensional Gaussian distribution around the deployment point.

In other words, when the deployment point of group G_i is at (x_i, y_i), we have the mean position $\mu = (x_i, y_i)$ and the pdf for node k in group G_i as

$$f_k^i(x, y|k \in G_i) = \frac{1}{2\pi\sigma^2} e^{-[(x-x_i)^2+(y-y_i)^2]/2\sigma^2}, \tag{6}$$

where σ denotes the standard deviation. Given such a group based Gaussian deployment model, we formulate the degree distribution of the key sharing graph based on both the deployment and the key distribution mechanism.

Let us consider any node location $H = (x, y)$ in the rectangular deployment field. Therefore, the probability that a node n_i from group i with deployment point (x_i, y_i) resides within the rectangular area $dxdy$ centered at point H is given by

$$\frac{1}{2\pi\sigma^2} \exp - \frac{(d_i H)^2}{2\sigma^2} \cdot dxdy$$

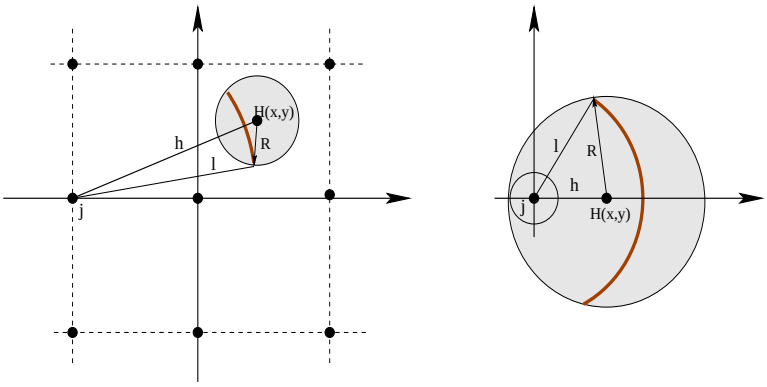


Fig. 1. Deployment points and resident point distribution

where $(d_iH)^2 = (x - x_i)^2 + (y - y_i)^2$. We denote

$$f(d_iH|n_i \in i) = \frac{1}{2\pi\sigma^2} e^{-\frac{(d_iH)^2}{2\sigma^2}} \tag{7}$$

and have the following result.

LEMMA 3.1. *If $g(x, y|j)$ denotes the probability that a node n_j from group j is within transmission radius R of point $H = (x, y)$, then $g(x, y|j) = \mathbf{1}\{h < R\}\left[1 - e^{-\frac{(R-h)^2}{2\sigma^2}}\right] + \int_{|h-R|}^{h+R} f(l|n_j \in j) \cdot 2lcos^{-1}\left(\frac{l^2+h^2-R^2}{2lh}\right) dl$, where $\mathbf{1}\{\cdot\}$ is the set indicator function.*

PROOF. When a sensor node resides at the point $H = (x, y)$ as shown in Fig. 1, the probability that the sensor node n_j from group j resides within the circle centered at location H with radius R is defined as $g(h|n_j \in \text{group}j)$, where $h = d_jH$, is the distance between H and the deployment point of group j. When $h > R$, as shown in the first diagram of Fig. 1,

$$g(x, y|j) = \int_{|h-R|}^{h+R} f(l|n_j \in j) \cdot 2lcos^{-1}\left(\frac{l^2 + h^2 - R^2}{2lh}\right) dl,$$

where the length of arc of the ring centered at j is calculated and then integrated over all possible values of l.

When $h < R$, as shown in the second diagram of Fig. 1,

$$g(x, y|j) = \int_0^{R-h} l \cdot 2\pi f(l|n_j \in j)dl + \int_{R-h}^{R+h} 2lcos^{-1}\left(\frac{l^2 + h^2 - R^2}{2lh}\right) f(l|n_j \in j)dl.$$

Thus,

$$g(x, y|j) = \mathbf{1}\{h < R\}\left[1 - e^{-\frac{(R-h)^2}{2\sigma^2}}\right] + \int_{|h-R|}^{h+R} f(l|n_j \in j) \cdot 2lcos^{-1}\left(\frac{l^2+h^2-R^2}{2lh}\right) dl,$$

where $\mathbf{1}\{\cdot\}$ is the set indicator function whose value is 1 when the evaluated condition is true and 0 otherwise, and $f(l|n_j \in j)$ is given by Eqn. 7. □

THEOREM 3.2. *The probability distribution of the degree of the key sharing topology, $p_g(k)$, assuming that the nodes are deployed in groups and reside according to a two dimensional gaussian distribution around the deployment points, is given by $p_g(k) = \int_0^Y \int_0^X \sum_{l=k}^N \binom{l}{k} q^k (1-q)^{l-k} N_b(l,x,y)$ where $N_b(l,x,y)$ is the probability that a node is at point (x,y) and it has l neighboring nodes.*

PROOF. Let $d = g(x,y)$ denote the probability that a node resides within a radius R of point (x,y). Then from Lemma 1, we get

$$d = g(x,y) = \sum_j g(x,y|j) Pr[j] \tag{8}$$

where $Pr[j]$ is the probability that a deployed node belongs to group j. We assume that a sensor node is selected to be in each group with an equal probability and is equal to $\frac{1}{tn}$.

Let $p(l|x,y)$ be the probability that there are l nodes within radius R of (x,y). Therefore,

$$p(l|x,y) = \binom{N}{l} d^l (1-d)^{N-l} \tag{9}$$

where N is the total number of nodes deployed. The probability that a deployed node is at point $H = (x,y)$, is given by

$$\sum_i f(d_i H | n_i \in i) \cdot Pr[i] dx dy$$

Let $N_b(l,x,y)$ be the probability that a node is at (x,y) and it has l neighboring nodes. Thus,

$$N_b(l,x,y) = p(l|x,y) \sum_i f(d_i H | n_i \in i) \cdot Pr[i] dx dy \tag{10}$$

Now, let $p_g(k|l,x,y)$ be the probability that a node which is located at (x,y) and has l neighbors shares keys with exactly k neighbors. Hence,

$$p_g(k|l,x,y) = \binom{l}{k} q^k (1-q)^{l-k} \tag{11}$$

where q is the key sharing probability. Therefore,

$$p_g(k,x,y) = \sum_{l=k}^{N} p_g(k|l,x,y) N_b(l,x,y) \tag{12}$$

Thus, integrating over the entire region, the degree distribution of a group based deployed network with each group deployed in a gaussian manner, is given by

$$p_g(k) = \int_0^Y \int_0^X p_g(k,x,y) = \int_0^Y \int_0^X \sum_{l=k}^{N} p_g(k|l,x,y) N_b(l,x,y) \tag{13}$$

□

Thus, based on both physical proximity and the probability of key sharing between neighbors, we get a degree distribution $p(k)$ for each of the graphs representing the two different deployment strategies. We will now perform our epidemic propagation analysis on these two types of random networks under the two scenarios of node recovery and

without node recovery. The random graph in our analysis is denoted by G, and $p_u(k)$ (for uniform deployment) and $p_g(k)$ (for group based deployment) characterize the degree distribution under the respective deployment strategies.

3.2 Network Connectivity

Before we consider our analysis of the epidemic processes on the overlay key sharing graph of the sensor network, it is essential to ensure that the graph is connected. We borrow the results from the works on connectivity in ad hoc networks presented in [Bettstetter 2002; Penrose 1999]. From their results, a geometric random graph with N nodes is k-connected with probability $P(k\text{-connected}) = \left(1 - \sum_{i=0}^{k-1} \frac{\left(\rho\pi R^2\right)^i}{i!} \cdot e^{-\rho\pi R^2}\right)^N$, where ρ is the network density and R is the transmission radius. Since we are considering 1-connectivity, the probability that the graph is connected is given by

$$P(connected) = \left(1 - e^{-\rho\pi R^2}\right)^N \tag{14}$$

Our goal is to make this probability almost equal to 1.

Without loss of generality on the deployment strategy, let $p(k)$ denote the degree distribution of the key sharing topology. Thus, $\delta = \sum_{k=1}^{N-1} kp(k)$ is the expected degree of the network. We observe that, in the expression for the aforementioned probability of connectivity of the network, $\rho\pi R^2$ is the expected number of neighbors that a node has. In other words, it can be interpreted as the expected number of neighboring nodes that fall within the transmission range of a given node. Thus, for our network with expected degree denoted by δ, the probability that the network is connected is given by

$$P(connected) = \left(1 - e^{-\delta}\right)^N \tag{15}$$

For our analysis, we consider the minimum value of the key sharing probability q to be such that it is *well above* the threshold in order to keep the network connected with very high probability.

3.3 Compromise Spread Without Node Recovery

Given the random graph construction based on the two deployment strategies, we now analyze the case of compromise spread when no node recovery is performed. In other words, a compromised sensor node will remain infectious indefinitely.

Let $G_0(x)$ be the generating function of the degree distribution $p(k)$ of a randomly chosen vertex in G and is defined by

$$G_0(x) = \sum_{k=0}^{\infty} p(k)x^k \tag{16}$$

The average degree z of G is denoted by

$$z = \sum_k kp(k) = G_0'(1) \tag{17}$$

Similarly, $G_1(x)$ denotes the degree distribution of the vertices at the end of randomly chosen edges. The distribution of degrees of vertices reached by following edges is pro-

portional to $kp(k)$ and thus the generating function for those degrees is

$$\frac{\sum_k kp(k)x^k}{\sum_k kp(k)} = \frac{xG_0'(x)}{G_0'(1)} \tag{18}$$

To elucidate further $G_1(x)$ represents the distribution of the number of ways of leaving these vertices excluding the edge we come along, which is the degree minus 1 and is given by

$$G_1(x) = \frac{1}{z}G_0'(x) \tag{19}$$

If λ denotes the infection probability of a node being infected by communicating with a compromised node, then the number c of compromised edges around a randomly chosen vertex is generated by

$$\begin{aligned} G_0(x;\lambda) &= \sum_{c=0}^{\infty}\sum_{k=c}^{\infty} p(k)\binom{k}{c}\lambda^c(1-\lambda)^{k-c}x^c \\ &= \sum_{k=0}^{\infty} p(k)\sum_{c=0}^{k}\binom{k}{c}(x\lambda)^c(1-\lambda)^{k-c} = \sum_{k=0}^{\infty} p(k)(1-\lambda+x\lambda)^k \\ &= G_0(1-\lambda+x\lambda). \end{aligned} \tag{20}$$

Similarly, $G_1(x;\lambda) = G_1(1-\lambda+x\lambda)$.

Let $H_1(x;\lambda)$ be the generating function that denotes the distribution of the sizes of the cluster of vertices or components reached by following a randomly chosen edge that is compromised. The degree of such an end vertex can vary from 0 to $N-1$. Moreover, if the degree is at least one, then following each edge out of that vertex would lead to more vertices whose degree distribution is also $H_1(x;\lambda)$. If there are k edges emanating from the vertex at the other end of the random edge, then the distribution of the sum of the sizes of the k clusters that each edge from the end vertex leads to, is given by $H_1(x;\lambda)^k$.

The generating function $H_1(x;\lambda)$ for the total number of nodes reachable or compromised as a result of a single transmission along an edge of the network is, thus, generated by a self consistency relation of the form [Newman et al. 2001; Newman 2002]

$$H_1(x;\lambda) = xG_1(H_1(x;\lambda);\lambda). \tag{21}$$

We are interested in the distribution of the size of the component to which a randomly chosen vertex belongs. In other words, the distribution of the number of nodes affected by an outbreak when the infection starts at a single infective node is generated by

$$H_0(x;\lambda) = x\sum_{k=0}^{N-1} p(k)[H_1(x;\lambda)]^k = xG_0(H_1(x;\lambda);\lambda). \tag{22}$$

The average size of the outbreak cluster is derived as $s = H_0'(1;\lambda)$ and is given by

$$s = 1 + \frac{\lambda G_0'(1)}{1-\lambda G_1'(1)}. \tag{23}$$

Infection probability λ essentially captures the spreading capability of the virus that could compromise the network: the larger it is, the stronger the virus is. We assume that its value can be obtained by means of measurement or analysis.

We remark here that in traditional epidemiology, the parameter λ, denoting the infection probability, generally represents the portion of the population that is susceptible to the infection. However, in a sensor network, it is typically assumed that all the nodes are homogeneous and therefore equally susceptible. Thus, in this case, instead of considering a fraction of the network as susceptible, we consider the whole network to be susceptible and subsequently, at t=0, all $N-1$ nodes are susceptible with one node being infected. The parameter λ, in this case, tries to capture the characteristic of the malware that is spreading and what technique it adopts, i.e., whether it has the properties of a worm, virus, or trojan, etc. In other words, using the variable λ, we try to parametrically capture the infectivity of a malware or the probability by which an infection spreads on a link between any pair of infected and susceptible node.

Thus, λ succeeds in differentiating between different malwares and their propagation characteristics and is assumed to be fixed for a particular spread but may vary from time to time based on the type of malware and what systemic technique it adopts to spread.

We, thus, use λ to connect the infection probability of the malware to the physical properties of the network (expressed in terms of p and q) and see if there is a resultant epidemic.

Given the above result, we can see that the outbreak point for the network is $\lambda = 1/G_1'(1)$ which marks the onset of an epidemic. For $\lambda > 1/G_1'(1)$ we have an epidemic in the form of a giant component in the random network. We observe that $H_0(1;\lambda)$ which is the distribution of the cluster formation is only valid below the threshold point beyond which it becomes invalid because in a giant component there could be loops and the recursive distribution of the end node degree of an edge as stated by Eqn. 21 would not hold. Thus, beyond the threshold point, we define $H_0(1;\lambda)$ to be the distribution of isolated clusters which do not have loop formations. If Ψ_m denotes the cluster size distribution of size m, then we observe that the fraction of the network forming the giant component is given by $S = 1 - \sum_m \Psi_m = 1 - H_0(1;\lambda)$.

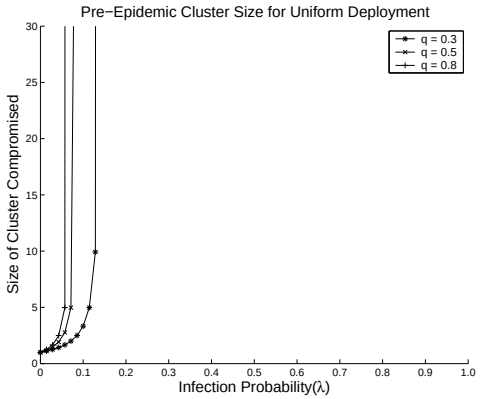
Rearranging and substituting from Eqn 22, we have $S = 1 - G_0(u;\lambda)$.

Here u is the root of the self-consistency relation

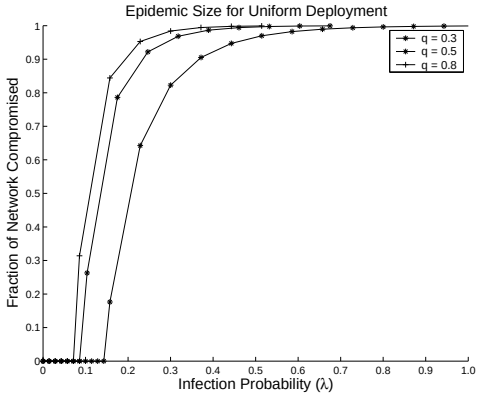
$$u = G_1(u;\lambda).$$

Intuitively, the above conclusion reveals that if $\lambda \leq 1/G_1'(1)$, the component of compromised nodes is finite in size regardless of the size of the network and each node's probability of being compromised is zero for large networks. On the contrary, if $\lambda > 1/G_1'(1)$, there always exists a finite probability for a node to be compromised.

We plot the effect of different key sharing probabilities on the epidemic outbreak on our two deployment strategies, viz uniform random and a collection of group based deployment strategies. Fig. 2 depicts this effect for a uniform random network with $N = 10000$ nodes deployed in a $600 \times 600 unit^2$ area with different key sharing probabilities q. The underlying physical topology is determined by the communication range of each node which is equal to 20 units. Given the physical deployment, we vary the probability of direct pairwise key sharing (q) and study the point of outbreak. As we can see in Fig. 2, while undoubtedly increasing q can facilitate communication in the network, the network also becomes more vulnerable to virus spreading. Specifically, when $q = 0.3$, network wide breakout is only possible when a compromised node has an infection probability (λ) larger than 0.15 to infect a neighbor. We note that in this case, we have an expected node degree of 15. On the contrary, λ only needs to be around 0.05 when $q = 0.8$ which subsequently



(a) Non-epidemic cluster size vs. infection probability (λ)

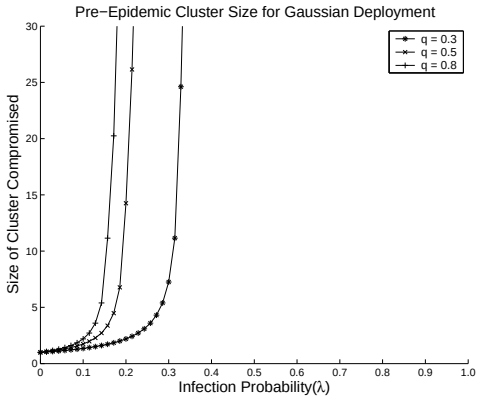


(b) Epidemic size vs. infection probability (λ)

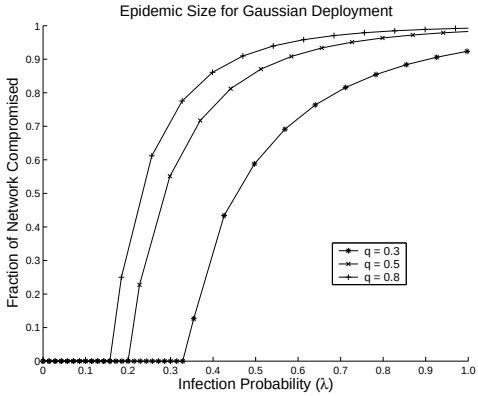
Fig. 2. Size of compromised node clusters for Uniform Deployment: (a) depicts the average size of infected clusters when there is no epidemic and (b) shows the epidemic size as the fraction of the entire network. The point where non-zero value appears indicates the transition from non-epidemic to epidemic

makes the expected node degree around 30. Fig. 2(b) illustrates the fraction of the network that is ultimately infected as the infection probability is increased beyond the critical point of the onset of outbreak. For instance, we observe that when key sharing probability is high ($q = 0.8$), the whole network is compromised with a λ value of around 0.4. On the contrary, with $q = 0.3$, the network could be compromised with only a high value of $\lambda = 0.7$. Although Fig. 2 (b) is a continuation of Fig. 2 (a) when the epidemic spreads to the entire network, a separate depiction provides a clear picture of the extent to which the network is infected before and after an epidemic. In summary, Fig. 2 clearly indicates the tradeoff between key sharing probability among sensor nodes and the vulnerability of the network to compromise.

In Fig. 3, we depict the same process with the deployment scheme changed to a group based one. The deployment points are arranged in a 10×10 grid with 100 nodes per group. The nodes in each group reside in a two-dimensional gaussian manner about their



(a) Non-epidemic cluster size vs. infection probability (λ)



(b) Epidemic size vs. infection probability (λ)

Fig. 3. Size of compromised node clusters for Gaussian Deployment: (a) depicts the average size of infected clusters when there is no epidemic and (b) shows the epidemic size as the fraction of the entire network. The point where non-zero value appears indicates the transition from non-epidemic to epidemic

mean deployment point with $\sigma = 10$. We observe that the potency of the propagation process is affected by the change in deployment. The λ values at which the epidemic starts to spread into the whole network has increased. For instance, for $q = 0.3$, the transition point is around $\lambda = 0.32$ as compared to be around 0.15 for the uniform deployment case. The reason for the decrease in the epidemic effect is caused by the fact that the expected node degree of the network is lowered when the nodes are deployed in groups distributed in a gaussian manner. This, obviously, has a crippling effect on the propagation process and thus helps in delaying the onset of the epidemic. However, this effect, as expected, slowly diminishes with increase in the variance of the gaussian distribution of each group, which gradually pushes the distribution to a more uniform nature. Thus, when we increase $sigma$ to around 40 in our deployment scenario with 10000 nodes, there is practically little difference between the two deployment strategies. This analysis shows that tuning the deployment parameters to a certain extent could result in making the network robust

against viral propagation without considerably hampering its connectivity.

3.4 Compromise Spread With Node Recovery

In this case, we assume that the network has the capability to recover some of the compromised nodes by either immunization or removal from the network. To capture this recovery effect, we assume that an infected node recovers or is removed from the network after an average duration of infectivity τ. In other words, a node in the sensor network remains infective for an average period τ after which it is immunized. During this infective period, the node transmits the epidemic to its neighbors with the infection rate β, denoting the probability of infection per unit time. Evidently, the parameter τ is critical to the analysis as it measures how soon a compromised node recovers. Naturally, we will perform our analysis following the SIR model in epidemic theory [Newman 2002].

First, consider a pair of adjacent nodes where one is infected and the other is susceptible. We define T as the compromise transmission probability, or in other words, the *transmissibility* of the infection. Given the above definitions for β and τ, we can say that the probability that the disease will not be transmitted from the infected to the susceptible is given by

$$1 - T = \lim_{\delta t \to 0} (1 - \beta \delta t)^{\tau/\delta t} = e^{-\beta\tau}. \tag{24}$$

Subsequently, we have the transmissibility

$$T = 1 - e^{-\beta\tau}.$$

In other words, the compromise propagation can be considered as a Poisson process, with average $\beta\tau$. The outcome of this process is the same as bond percolation and T is basically analogous to the bond occupation probability on the graph representing the key sharing network. Thus, the outbreak size would be precisely the size of the cluster of vertices that can be reached from the initial vertex (infected node) by traversing only occupied edges which are occupied with probability T. Notice that T explicitly captures node recovery in terms of the parameter τ.

Replacing λ with T in Equation 23, and following similar steps, we get the size of the average cluster as

$$s = 1 + \frac{TG_0'(1)}{1 - TG_1'(1)}. \tag{25}$$

and the epidemic size is obtained by

$$S = 1 - G_0(u; T). \tag{26}$$

where u is obtained by

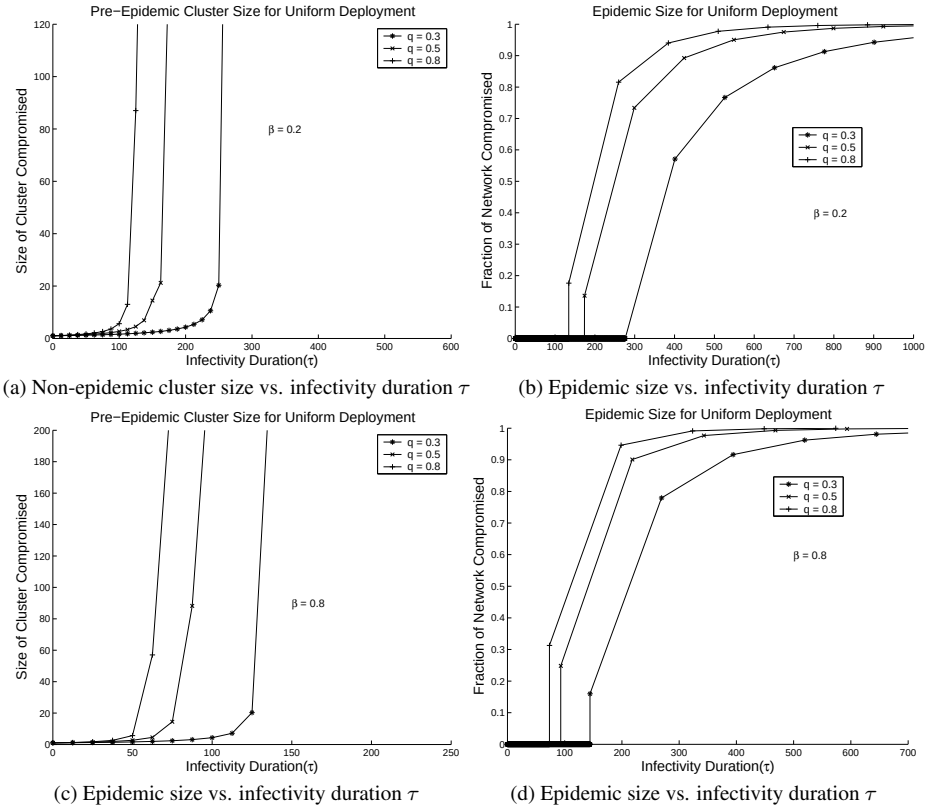
$$u = 1 - G_1(u; T), \tag{27}$$

and $G_0(u; T)$ and $G_1(u; T)$ are given respectively by

$$G_0(u; T) = G_0(1 + (u - 1)T), \tag{28}$$

and

$$G_1(u; T) = G_1(1 + (u - 1)T). \tag{29}$$



(a) Non-epidemic cluster size vs. infectivity duration τ

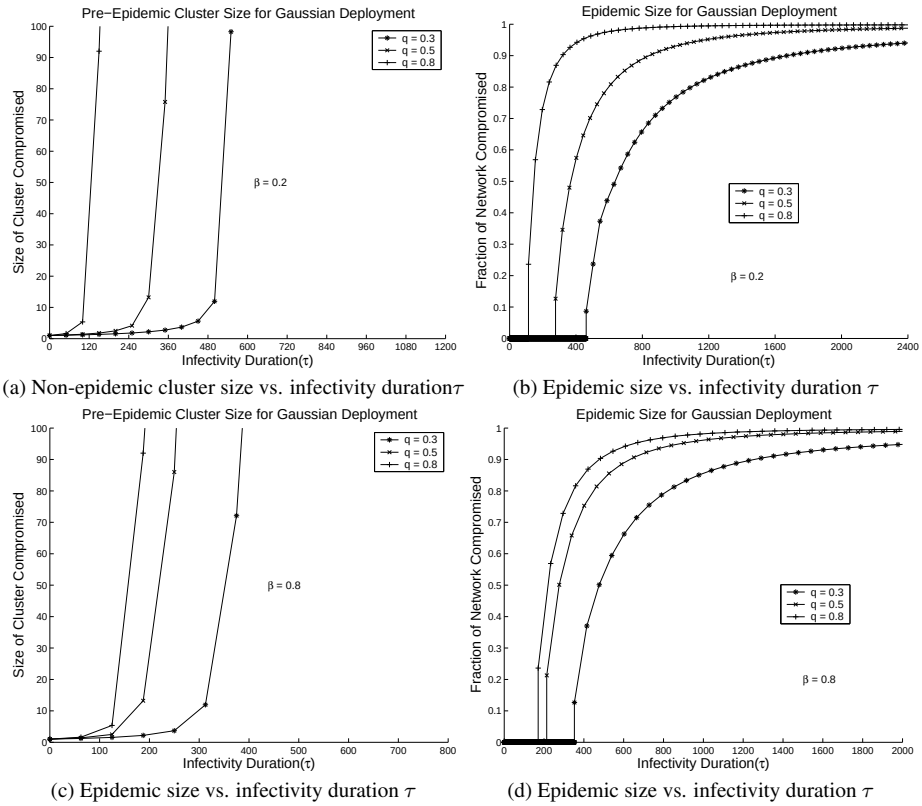
(b) Epidemic size vs. infectivity duration τ

(c) Epidemic size vs. infectivity duration τ

(d) Epidemic size vs. infectivity duration τ

Fig. 4. Extent of Epidemic Size with Varying Infectivity Duration (Uniform Deployment): (a) Pre-Epidemic Cluster Size with Low Infection Probability (b) Post-Epidemic Infected Fraction with Low Infection Probability (c) Pre-Epidemic Cluster Size with High Infection Probability (d) Post-Epidemic Infected Fraction with High Infection Probability

Figs. 4 and 5 summarize this effect for the uniform and group based deployment respectively. They depict the epidemic outbreak against the average recovery time τ for low and high infection rates $\beta = 0.2$ and $\beta = 0.8$. The network setup is the same as before with $N = 10000$ and a 10×10 grid for the group deployment scheme. For uniform deployment, Figs. 4(a) and (b) depict the pre-epidemic and post-epidemic scenario when the infection rate is low ($\beta = 0.2$). In Figs. 4(c) and (d), the infection rate is high ($\beta = 0.8$). The plots are for different values of the key sharing probability q which governs the connectivity of the key sharing sensor topology. Fig. 5 shows the plots when the nodes are deployed in groups. Comparing the plots in Figs. 4 and 5 we observe similar characteristics as observed in the non-recovery case. Here, in the group based plots, the compromise process attains epidemic proportions at higher values of the infectivity duration τ. For instance, comparing Fig. 4(d) and Fig. 5(d), when $q = 0.3$ and $\beta = 0.8$, the epidemic outbreak for the group deployment starts at around $\tau = 500$, whereas in the uniform deployment scenario, its onset is when τ is around 150. This indicates that the potency of the compromise is reduced by the gaussian distribution deployment model. In other words, the nodes in the group based deployment model would need to stay infective for a larger average dura-



(a) Non-epidemic cluster size vs. infectivity durationτ

(b) Epidemic size vs. infectivity duration τ

(c) Epidemic size vs. infectivity duration τ

(d) Epidemic size vs. infectivity duration τ

Fig. 5. Extent of Epidemic Size with Varying Infectivity Duration (Gaussian Deployment): (a) Pre-Epidemic Cluster Size with Low Infection Probability (b) Post-Epidemic Infected Fraction with Low Infection Probability (c) Pre-Epidemic Cluster Size with High Infection Probability (d) Post-Epidemic Infected Fraction with High Infection Probability

tion in order to spread to the entire network than when uniformly deployed. The expected duration of a node's infectivity could be a possible measure of the degree of potency of the viral infection and this comparison is indicative of which deployment scheme is better poised to resist against an outbreak.

In next section we will discuss about an important parameter in epidemic theory whose value indicates whether an infection can potentially result in an epidemic. We derive an expression for this parameter for our sensor network model and investigate its behavior.

4. BASIC REPRODUCTIVE NUMBER

In epidemiology, the *Basic Reproductive Number* R_0 is defined as the expected number of people that a single infective individual can infect in a pool of mostly susceptible candidates. Its importance lies in the fact that it characterizes the epidemic growth at the start of an outbreak: the infection will eventually die out when $R_0 < 1$ and when $R_0 > 1$, the disease will spread exponentially and consequently may lead to a large epidemic.

Since we have proposed an epidemic model for the spread of node compromise, it is essential that we verify how our network parameters contribute to the derivation of the impor-

tant epidemic parameter R_0. We know that the epidemic threshold for R_0 is 1. Therefore, expressing R_0 in terms of the relevant network parameters like the key sharing probability q, the infection rate β, the infectivity duration τ, etc., would shed more light on what parameter to control in order to prevent an epidemic outbreak.

The resultant expression of the basic reproductive number would also be indicative of the correctness of our epidemic model.

THEOREM 4.1. *If T denotes the transmissibility and u denotes the probability that the vertex at the end of a randomly chosen edge remains uninfected during an epidemic, then the basic reproductive number R_0 is given by*
$R_0 = \sum_j j \sum_k \binom{k}{j} T^j (1-T)^{k-j} \frac{(1-v^k)p(k)}{\sum_k [(1-v^k)p(k)]}$, *where* $v = 1 - T + Tu$.

PROOF. Let I denote the set of infected nodes in the sensor network. Moreover, let n_i denote the number of nodes that are infected by node i and d_i represent the degree of node i. We are interested in the probability $Pr[n_i = j | i \in I]$. This can be written as

$$Pr[n_i = j | i \in I] = \sum_k Pr[n_i = j | d_i = k, i \in I] Pr[d_i = k | i \in I] \tag{30}$$

From Equation (30), using Bayes' Rule, we can write

$$Pr[d_i = k | i \in I] = \frac{Pr[i \in I | d_i = k] p(k)}{Pr[i \in I]}. \tag{31}$$

where $p(k)$ is the degree distribution of the network. Given that β is the infection probability per unit time and τ is the average recovery time of an infected node, we have from Equation (24) the probability with which each link is occupied as

$$T = 1 - e^{-\beta\tau}. \tag{32}$$

From Equation (26) and (27), we have the size of the epidemic as S and u is simply the probability that the vertex at the end of a randomly chosen edge remains uninfected during an epidemic. Thus, the probability that a vertex does not become infected via one of its edges is

$$v = 1 - T + Tu, \tag{33}$$

where $1 - T$ is the probability that the edge is unoccupied, and Tu denotes that probability that it is occupied but connects to an uninfected vertex. Consequently, the total probability of being *uninfected* if a vertex has degree k is v^k. This leads to

$$Pr[i \in I | d_i = k] = 1 - v^k. \tag{34}$$

Hence, we have

$$Pr[i \in I] = \sum_k (1 - v^k) p(k) \tag{35}$$

Substituting this into Equation (31) gives

$$Pr[d_i = k | i \in I] = \frac{(1 - v^k) p(k)}{\sum_k (1 - v^k) p(k)}. \tag{36}$$

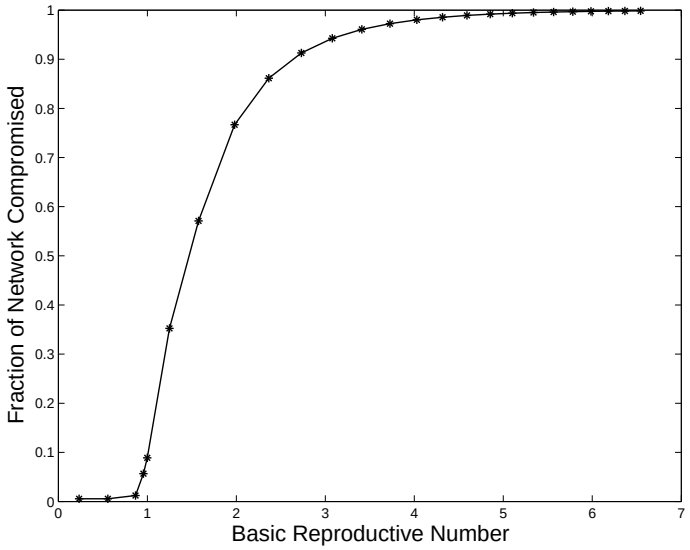


Fig. 6. Fraction of network infected vs. basic reproductive number R_0

Furthermore, we have

$$Pr[n_i = j | d_i = k, i \in I] = \binom{k}{j} T^j (1-T)^{k-j} \tag{37}$$

and substituting this into Equation (30) gives

$$Pr[n_i = j | i \in I] \; = \; \sum_k \binom{k}{j} T^j (1-T)^{k-j} \frac{(1-v^k)p(k)}{\sum_k [(1-v^k)p(k)]}. \tag{38}$$

Finally, the basic reproductive number R_0 is given by

$$R_0 = \sum_j j Pr[n_i = j | i \in I]. \tag{39}$$

□

Equation (39) gives an expression where R_0 is expressed in terms of the transmissibility T which is dependent on two factors, namely, the infection rate β and the infective duration τ. Fig. 5 illustrates the epidemic size based on the basic reproductive number R_0. As expected, we find that in our random graph based sensor network model, the transition point of R_0 is 1, above which an epidemic outbreak occurs. This result further proves the correctness of the model for capturing the epidemic propagation of a malware infection in a securely communicating sensor network.

5. SIMULATION

We employ a discrete-event dynamic system and therefore event-driven simulation to accurately simulate the propagation of the infection spreading process. We have used JProwler

[jpr], a probabilistic, event-driven wireless network simulator in Java, for our experiments. JProwler, which is a java implementation of the Prowler simulator is capable of simulating the non-deterministic nature of the communication channel and the low-level communication protocol of wireless sensor nodes. A probabilistic radio channel model is used with the received signal strength function defined as

$$P_{rx} = \frac{P_{tx}}{1 + d^{\gamma}} \cdot [1 + \alpha(d)] \cdot [1 + \beta(t)] \tag{40}$$

where P_{tx} is the transmit power and d is the distance between two nodes. The parameters α and β are normal random variables and model the probabilistic nature of the radio channel. A packet error rate p_{error} simulates the effect of any unmodeled effects on the transmission probability. In this section, we outline our discrete-event driven simulation model setup for the gradual progress of the spread of node compromise. We then use this model to capture the time dynamics of the spread of the compromise which we have largely omitted in our random graph based epidemic model. In our random graph analytical model, we obtained the static values of the maximum fraction of the network that was compromised. Through our simulation, we aim to capture the dynamics of the infection spread. This way, we not only concern the final stable state results but also investigate the temporal effects of node recovery on the extent of infection spread.

5.1 Simulation Setup

In our simulation, we assume the number of sensor nodes in the network to be 10000. The sensor network is produced by distributing the sensors in a 600×600 $unit^2$ area. The communication range of each node is assumed to be 25 units. The mean data rate of the wireless links is set to 40 Kbps with the packet length set to 48 bytes. The mean packet transmission time is calculated accordingly. The MAC layer is a simple CSMA based scheme modeling the Berkeley motes' MAC layer. We have used the default settings of JProwler for the standard deviations of α and β as 0.45 and 0.02, respectively and the packet loss rate is set to 0.1. Each point in our simulation is the result of 30 runs and is depicted as the average of these 30 runs with a 95% confidence interval.

For the uniform random deployment, the location of each node is selected from a uniform distribution within the region. For the group based gaussian deployment scenario, the deployment points are arranged in a 10×10 grid in the monitored area and there are 100 nodes in each group. For each group, the location of a node is selected from a two-dimensional gaussian distribution with the deployment point as the mean and standard deviation $\sigma = 10$.

We employ the random key pre-distribution scheme described in [Eschenauer and Gligor 2002] to establish the pairwise shared keys among sensor nodes. For each pair of neighbors, a small subset of keys are chosen randomly from a large pool such that they share at least one key with probability q.

Our simulation works in two phases. In the first phase, we form the network where each node identifies its set of neighbors and entries are made into a neighbor table. Based on typical communication distances between nodes and their respective locations, we derive the set of neighbors for each node. The degree of the key sharing network is controlled by changing the value of the key sharing probability q between neighbors. It is in the first phase that the key sharing topology for the epidemic propagation is derived based on the value of q between each pair of neighboring nodes and from the manner in which the nodes

are deployed.

In the second phase, we simulate actual virus propagation. Initially, at $t = 0$, the number of infected nodes, denoted by $I(0)$ is set to be 1. At any time point t, the population is divided into the group of susceptible nodes, $S(t)$, and the group of infected nodes, $I(t)$. In the situation where we have nodes that are immunized and thus recovered, we denote this set of recovered nodes by $R(t)$.

The timeline of these sub-populations is obtained by observing the population counts after fixed simulation intervals of 1 time unit. The average incubation time at an infected node is assumed to follow a poisson distribution with a considerably low mean of 1 unit. This denotes the time period for a node to evolve from an infected state to an infective state. The low average value essentially dictates that an infected node at simulation time t will be ready to infect its neighbors at time $(t + 1)$ with a high probability. Furthermore, when this incubation period is over, we assume that the time it takes for the infected node to infect its susceptible neighbor is negative exponentially distributed with a mean of 1 unit time.

There are two simulation scenarios corresponding to our analysis.

5.1.1 *No Recovery.* First we perform the simulation for the case where nodes once compromised are not recovered. Here, the simulation is based primarily on one event - the *Infection Event* which is an infected packet transmitted from an infected node to its susceptible neighbor. Associated with each Infection event packet are the node ID of the source and the ID of the destination node that has been infected by this source. The seed for the simulation is a single Infection event with a randomly selected node ID from among its neighbors as the infected destination and the source as null. The simulation is started by inserting this event into the event priority queue with the prioritization based on the event times. When an *Infection Event* is popped from the queue, the list of its neighbors are looked up in the neighbor table. From its susceptible neighbors, a node is selected randomly for infection, according to the infection probability β and an Infected packet is transmitted to it. A packet transmission event takes into consideration the physical characteristics of the network like transmission time and packet collision rate. Upon being infected, a node generates a new infection event.

5.1.2 *With Recovery.* In the case where infected nodes are recovered, we define a new *Recovery Event* for our simulation model. Our aim is to keep the average recovery time constant and study the time dynamics of the nodes in different groups based on different topology structure and infection probabilities. We assume the recovery time for an infected node to be negative exponentially distributed with a mean of τ_0 units. The CDF of the recovery time is represented by

$$Pr[t < T] = 1 - e^{-\tau_0 T}. \tag{41}$$

When an *Infection Event* is popped from the queue, we obtain the difference of the current simulation time and the time when the event was inserted into the queue. Using this time difference in Equation (41), the probability of inserting a *Recovery Event* or an *Infection Event* is calculated. However, when a *Recovery Event* is triggered, no event is further inserted. The corresponding node is marked as recovered and remains immune to further infections.

5.2 Simulation Results and Discussion

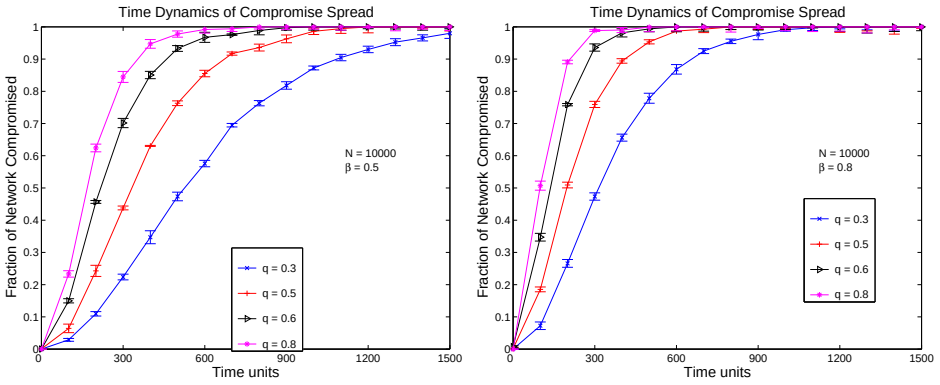


Fig. 7. Dynamics of the infective population (Uniform Random Deployment Without Node Recovery) (a) Moderate Infectivity, (b) High Infectivity

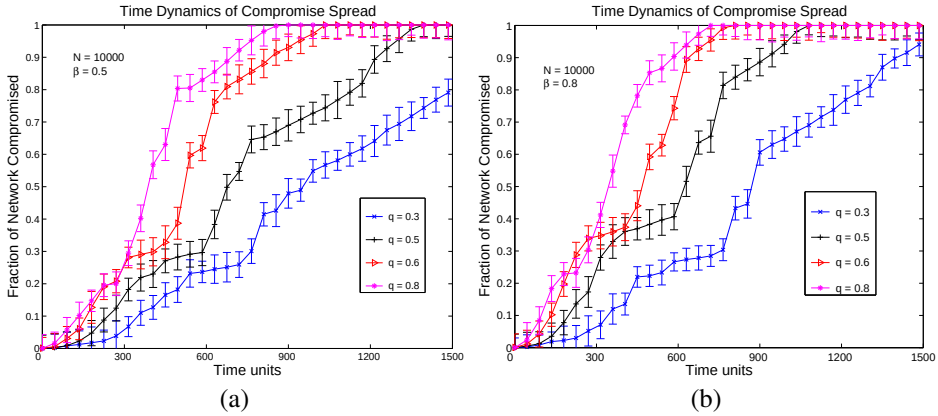


Fig. 8. Dynamics of the infective population (Group Based Deployment Without Node Recovery) (a) Moderate Infectivity, (b) High Infectivity

5.2.1 *Simulation Results for No Recovery Case.* The simulation results for the case without recovery are shown in Figs. 7 and 8. Fig. 7 represents the case for uniform random deployment while Fig. 8 shows the case for the group based deployment. The figures illustrate the compromise dynamics under moderate and high infectivity β. We vary the key sharing probability q between neighbors in each plot in order to simulate the variance of the degree distribution of the key sharing topology under the two deployment scenarios. As expected, the uniform random deployment curves are smoother than the group based deployment. In the latter, the slope is sharply affected by the density of nodes in the region of the propagation spread. This varies regularly and we observe that this variation of node density actually slows down the propagation dynamics. For instance, comparing Fig. 8(b) with Fig. 7(b), we observe that for $q = 0.5$, the network is compromised around time $t = 1400$ for the group deployment case, whereas for uniform deployment, the network is

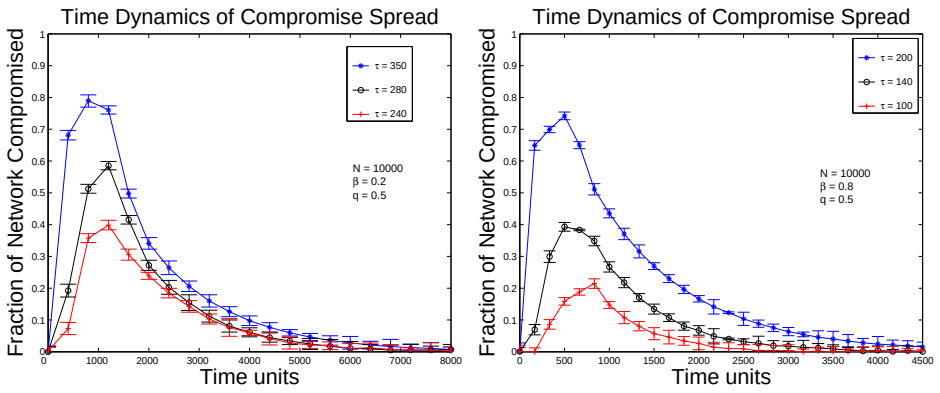


Fig. 9. Dynamics of the infective population for Uniform Random Deployment (With Node Recovery and $q = 0.5$) (a) Low Infectivity, (b) High Infectivity

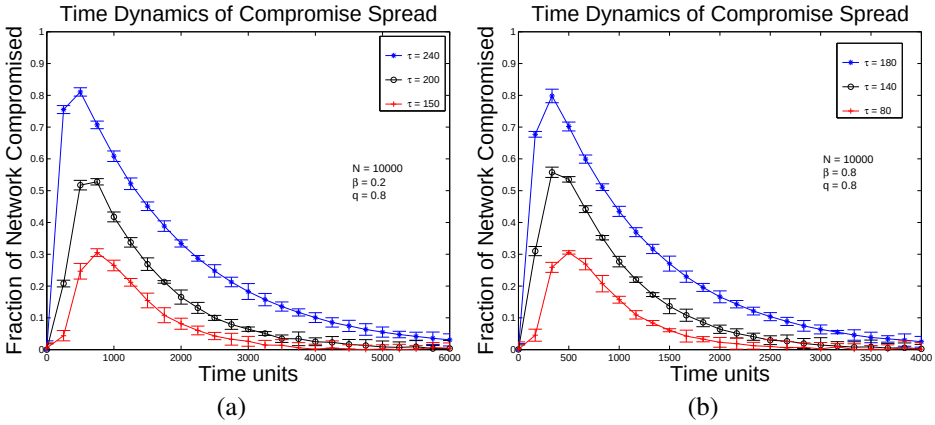
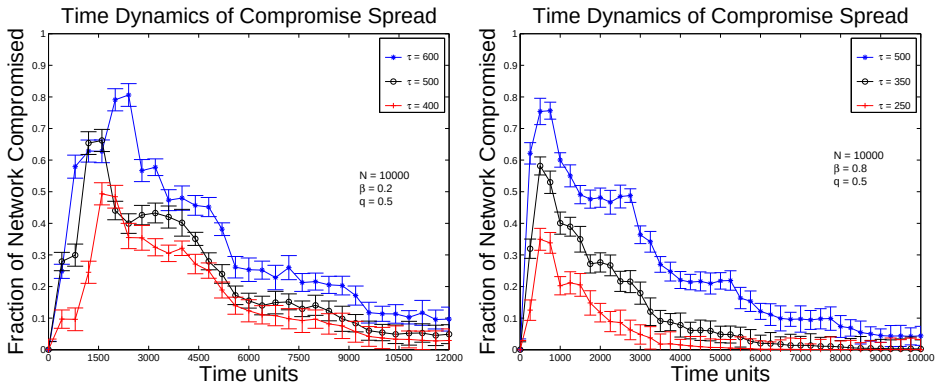


Fig. 10. Dynamics of the infective population for Uniform Random Deployment (With Node Recovery and $q = 0.8$) (a) Low Infectivity, (b) High Infectivity

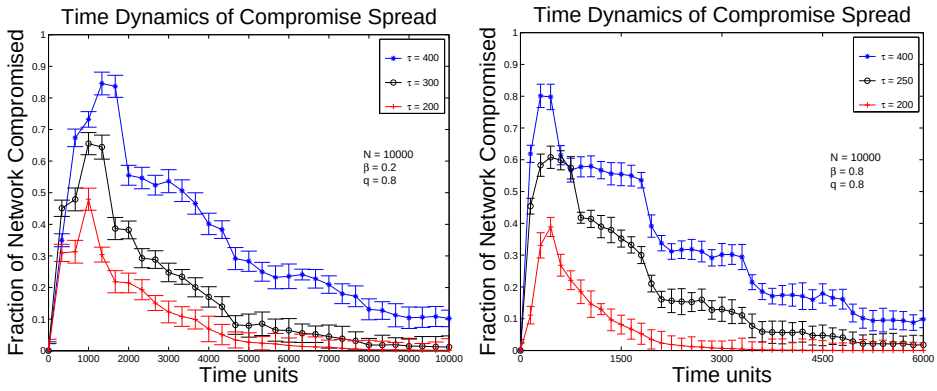
compromised around time $t = 900$ for the same value of q. A similar observation is made when comparing Figs. 7(a) and 8(a) for a lower infectivity value of $\beta = 0.5$. The reason for the difference in time taken for the infection to spread to the entire network is attributed to the difference of the expected node degree in the different deployment cases. A group based deployment with each group deployed in a two dimensional gaussian manner results in a lowering of the expected node degree of the network at the physical level.

5.2.2 *Simulation Results for Recovery Case.* Figs. 9, 10, 11, and 12 show the simulation dynamics in the presence of a recovery strategy. Figs. 9 and 10 depict the epidemic propagation under a uniform random deployment of sensors, while Figs. 11 and 12 are depictions of a group based deployment with the sensors in each group distributed in a two-dimensional gaussian manner. For both types of deployment, we investigate two



(a) (b)

Fig. 11. Dynamics of the infective population for Group based deployment (With Node Recovery and $q = 0.5$) (a) Low Infectivity, (b) High Infectivity



(a) (b)

Fig. 12. Dynamics of the infective population for Group based deployment (With Node Recovery and $q = 0.8$) (a) Low Infectivity, (b) High Infectivity

scenarios, viz., one in which the infectivity of the compromise process is low and the second where it is quite high. For each of the cases of infectivity, we also observe the dynamics under different key sharing probabilities q. For a comparatively sparser key sharing network ($q = 0.5$), as depicted in Figs. 9(a) and (b), we observe that a higher infectivity duration is required to achieve similar levels of epidemic propagation in the network as in Fig. 10. Comparing Figs. 9 (a) and 10 (a), we observe that increasing the node connectivity by increasing q from 0.5 to 0.8, raises the infected population peak count from 40% to 80% for $\tau = 240$.

Trivially, we also observe that the peak count is achieved at a much faster rate with an increased key sharing probability from $q = 0.5$ to $q = 0.8$.

The plots in Figs. 11 and 12 depict the same dynamics of epidemic propagation, but here the nodes are deployed in groups which are gaussian distributed about the deployment

point. In Fig 11, each group is composed of 10 nodes, whereas in Fig 12, there are 50 nodes in each group. As in the case for no-recovery, we observe that the temporal behavior of the infective population reflects the variability of the node density at different regions of the network. As a result, the slope of the infective curve changes depending on how close to the mean position the propagation front of the compromise process is. In other words, we observe that when the curve is rising, the slope decreases when the compromise process wavefront reaches a sparse section of the network, i.e., an area farther away from any deployment points. On the other hand, when the curve is decreasing, it falls steeply when the wavefront reaches a sparse region because nodes are not infected as fast as they are recovering resulting in a net decrease in infected population. This is the reason we also observe an increase in the total infective percentage at certain phases of the process when the region becomes very dense. In such a situation, the infection process suddenly accelerates resulting in a net increase in the infective population percentage. Figs. 11 (a) and (b) depict the process in a network of 10000 nodes under two levels of infectivity β when the key sharing probability is 0.5, while Figs. 12 (a) and (b) depict the same for a higher key sharing probability of $q = 0.8$. Apart from the periodic increase and decrease in the propagation process for the gaussian distribution, we also observe an important comparative result between the two types of deployment.

Comparing Figs. 9 and 10 with Figs. 11 and 12, we notice that similar peak levels of infectivity are obtained for the group deployment at higher levels of the infectivity duration τ. In other words, the nodes have to remain infective for a longer duration in the group based deployment scenario than when uniformly deployed in order to infect the same fraction of the population. This is in tandem with what we observed in our previous analytical derivations and simulations where epidemic proportions are reached for larger values of the infectivity duration τ, than in the situation when nodes are uniformly deployed. As mentioned earlier, the reason for which we observe this difference is that the average node degree of the network is reduced when the nodes are deployed in groups. The propagation progresses much more smoothly when the connectivity is uniform in nature than when it varies every now and then. In our simulations, we observed that the compromise process is more adversely affected by regions of low connectivity than helped by regions of higher connectivity. If the process dies out in a region of low density before it could reach a region of higher density, then the nodes in these sparse areas become regions of failure for the epidemic process.

From the plots we observe that the average variance of the curves are higher in the group based deployment scheme than the uniform random delployment one. This is an expected observation because in the different runs for the gaussian deployment scheme, the network topology varied more than the uniform case. Although the deployment points for each group was the same in all the cases, the ultimate resident points varied, resulting in different connectivities and a resultant higher value for the variance around the mean values of the infected population.

We also find the average variance increased in the case of a simultaneous recovery than when there was no recovery. We attribute this change to the fact that the execution of two processes of infection and recovery simultaneously, as opposed to just one infection process, results in a slightly increased average variance for the recovery case.

5.3 Correlation between Analytical and Simulation Results

Although our analysis and simulation provide separate viewpoints of the epidemic process, they are correlated and a few observations connecting them reinforce the results. As mentioned earlier, in our random graph analysis, we obtained a final picture of the resulting epidemic fallout. However, our simulations captured the temporal dynamics of how the ultimate results were obtained and how the process evolved.

In this subsection, we observe some of the correlative aspects of our analytical and simulation results by focusing on the points of the analytical curves that are represented in our simulation results. For instance, we closely observe the peak values of the curves in the simulation results for the recovery case of both the uniform and group based deployment scenarios. These points represent the maximum fraction of the network compromised before the recovery process caused the network to recover. These points are represented in the analytical plots because the points of the curves represent the ultimate fraction of the network compromised given the values of the parameters β and q. We find that the values are close to each other. For instance, if we compare Fig. 11(b) with Fig. 5(d), we observe that the peak values of the simulation curves in Fig. 11(b), are very close to the points in the analytical curve for $q = 0.5$ in Fig. 5(d), with corresponding τ values. Similarly, for the analytical curve with $q = 0.8$ in Fig. 5(d), we observe that the peak infective values match closely with the simulation curves in Fig 12(b). Close correlation is also observed for other pairs of simulation and analytical curves.

From our analysis and simulations, we therefore remark that both the analytical and experimental results have significant implication for security scheme design in terms of revoking/immunizing compromised nodes in wireless sensor networks. While the simulation results dictate the speed at which the network must react in order to contain/prevent the effect of network wide epidemic, the analytical plots indicate what values of the key sharing probability should be, in a securely communicating network using private keys, in order to contain an infection spread below the epidemic threshold while still maintain connectivity to promote network-wide communications.

6. RELATED WORK

The mathematical modeling of epidemics is well documented [Anderson and May 1992; May and Lloyd 2001; Hethcote 2000; Bailey 1975]. In fact, visualizing the population as a complex network of interacting individuals has resulted in the analysis of epidemics from a network or graph theoretic point of view[Moore and Newman 2000; Pastor-Satorras and Vespignani 2001b; 2001a]. Specifically, the scale free topology has been of keen interest [Pastor-Satorras and Vespignani 2001b; Barthlemy et al. 2004; Newman 2002] and this model has been the basis for the analysis and extensive study of virus and worm spreading in the Internet [Staniford et al. 2002; Kephart et al. 1993; Kephart and White 1993].

Node compromise in sensor networks and the need for their security has also received immense attention [Alarifi and Du 2006]. A large portion of current research on security in sensor networks has been focused on protocols and schemes for securing the communication between nodes [Liu and Ning 2003; Eschenauer and Gligor 2002; Malan et al. 2004]. In [Eschenauer and Gligor 2002], the authors propose a random key distribution scheme for secure communication among sensor nodes. In [Liu and Ning 2003], the authors improve on the work in [Eschenauer and Gligor 2002] by taking advantage of node location information to improve key connectivity. Critical thresholds on connectivity of the random

graph induced by the random key predistribution scheme are investigated by the authors in [Yagan and Makowski 2008]. In [Du et al. 2004], the authors discuss a key management scheme based on node deployment knowledge. They consider a group based deployment where the resident points of nodes in each group follow a two-dimensional gaussian distribution around the deployment point of the group. In [Pietro et al. 2006], the authors provide critical values of the size of the keyring and the key pool such that the network is not only connected but also resilient against the capture of a fixed fraction of the nodes and their keys. However, we consider a dynamic process whereby the adversary acquires more keys by propagating the node compromise process from a small set of nodes. Revocation of keys of compromised nodes has been studied in [Chan et al. 2005].

In [Alarifi and Du 2006], the authors assert the importance of physical compromise of sensor nodes and propose an obfuscation and diversification mechanism to protect the secret keys of the nodes. Unfortunately, little work has been done on the defense strategies when the compromise of a single node could be used to compromise other nodes over the air. Our work takes the first step towards modeling this potentially disastrous propagation[De et al. 2006]. Connectivity issues in random ad hoc networks are extremely important as a pre-requisite before any epidemic-like propagation process is analyzed. In [Bettstetter 2002; Penrose 1999], the authors derive threshold values of the transmission range of the nodes that ultimately make the network k-connected with a given probability. The thresholds for the monotone properties of random geometric graphs have also been dealt with in [Goel et al. 2004]. We adopted some of the results presented in [Newman 2002] where the author proposes a percolation theory based evaluation of the spread of an epidemic on graphs with given degree distributions. However, their work is a generic analysis of epidemics in random graphs. In our work, we have considered the specific characteristics of sensor networks including distance, deployment and key constrained communication patterns. Furthermore, little has been shown there on the temporal dynamics of the epidemic spread and only final outcomes of an infection spread in a network is studied.

7. CONCLUSION

In this paper, we investigate the potential threat for compromise propagation in wireless sensor networks. Based on the principles of epidemic theory, we model the process of compromise spreading from a single compromised node to the whole network. In particular, we focus on the effects of the key factors of the network determining a potential epidemic outbreak where the whole network will be affected. Due to the unique distance and key sharing constrained communication pattern, we resort to a random graph model which is precisely generated according to the parameters of the real sensor network and perform the study on the graph. We also ensure that the key sharing network generated is connected before performing our epidemic propagation analyses. We also introduce the effect of node recovery after compromise and adapt our model to accommodate this effect. Moreover, we perform a comparative study of the effect of two deployment strategies on the outcome of the epidemic propagation. Our results indicate that a uniform random deployment is more vulnerable to epidemic propagation than a group based deployment model and reveal key parameters of the network in defending and containing potential epidemics. In particular, with node recovery, the result provides benchmark time period for the network to recover a node in order to defend against the epidemic spreading and also critical values of the key sharing probability which characterize the transition from a non-epidemic to an epidemic

state of the network compromise. Our extensive simulation results validate our analytical results and more importantly, provide insights into the dynamics of the system in terms of temporal evolution of the infection process.

Acknowledgements : We are grateful to the anonymous referees for their constructive comments which helped us significantly improve the technical content of the paper.
This work is supported by NSF ITR grant under Award Number IIS-0326505 and Texas Advanced Research Program under grant No. 14-748779.

REFERENCES

JProwler. http://www.isis.vanderbilt.edu/projects/nest/jprowler/.

AKYILDIZ, I. F., SU, W., SANKARASUBRAMANIAM, Y., AND CAYIRCI, E. 2002. A survey on sensor networks. *Communications Magazine, IEEE 40,* 8, 102–114.

ALARIFI, A. AND DU, W. 2006. Diversify sensor nodes to improve resilience against node compromise. In *SASN '06: Proceedings of the fourth ACM workshop on Security of ad hoc and sensor networks*, New York, NY, USA, pp. 101–112. ACM.

ANDERSON, R. M. AND MAY, R. M. 1992. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press.

BAILEY, N. 1975. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

BARTHLEMY, M., BARRAT, A., PASTOR-SATORRAS, R., AND VESPIGNANI, A. 2004. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Phys. Rev. Lett. 92*, 178701.

BETTSTETTER, C. 2002. On the minimum node degree and connectivity of a wireless multihop network. In *MobiHoc*, pp. 80–91. ACM.

BRAGINSKY, D. AND ESTRIN, D. 2002. Rumor routing algorthim for sensor networks. In C. S. RAGHAVENDRA AND K. M. SIVALINGAM (Eds.), *WSNA*, pp. 22–31. ACM.

CALLAWAY, D. S., NEWMAN, M. E. J., STROGATZ, S. H., AND WATTS, D. J. 2000. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters 85*, 5468.

CHADHA, A., LIU, Y., AND DAS, S. K. 2005. Group key distribution via local collaboration in wireless sensor networks. In *IEEE International Conference on Sensor and Ad Hoc Communications and Networks (SECON)*. IEEE Computer Society.

CHAN, H., GLIGOR, V. D., PERRIG, A., AND MURALIDHARAN, G. 2005. On the distribution and revocation of cryptographic keys in sensor networks. *IEEE Trans. Dependable Sec. Comput. 2,* 3, 233–247.

CHAN, H., PERRIG, A., AND SONG, D. X. 2003. Random key predistribution schemes for sensor networks. In *IEEE Symposium on Security and Privacy*, pp. 197–213. IEEE Computer Society.

CHONG, C.-Y. AND KUMAR, S. P. 2003. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE 91,* 8, 1247–1256.

DE, P., LIU, Y., AND DAS, S. K. 2006. Modeling node compromise spread in wireless sensor networks using epidemic theory. In *WOWMOM*, pp. 237–243. IEEE Computer Society.

DU, W., DENG, J., HAN, Y. S., CHEN, S., AND VARSHNEY, P. K. 2004. A key management scheme for wireless sensor networks using deployment knowledge. In *INFOCOM*, pp. 597.

DU, W., DENG, J., HAN, Y. S., VARSHNEY, P. K., KATZ, J., AND KHALILI, A. 2005. A pairwise key predistribution scheme for wireless sensor networks. *ACM Trans. Inf. Syst. Secur. 8,* 2, 228–258.

ESCHENAUER, L. AND GLIGOR, V. D. 2002. A key-management scheme for distributed sensor networks. In V. ATLURI (Ed.), *ACM Conference on Computer and Communications Security*, pp. 41–47. ACM.

GOEL, A., RAI, S., AND KRISHNAMACHARI, B. 2004. Sharp thresholds for monotone properties in random geometric graphs. In L. BABAI (Ed.), *STOC*, pp. 580–586. ACM.

HETHCOTE, H. W. 2000. The mathematics of infectious diseases. *SIAM Rev. 42,* 4, 599–653.

HUI, J. W. AND CULLER, D. E. 2004. The dynamic behavior of a data dissemination protocol for network programming at scale. In J. A. STANKOVIC, A. ARORA, AND R. GOVINDAN (Eds.), *SenSys*, pp. 81–94. ACM.

KEPHART, J. O. AND WHITE, S. R. 1993. Measuring and modeling computer virus prevalence. In *SP '93: Proceedings of the 1993 IEEE Symposium on Security and Privacy*, Washington, DC, USA, pp. 2. IEEE Computer Society.

KEPHART, J. O., WHITE, S. R., AND CHESS, D. M. 1993. Computers and epidemiology. *IEEE Spectr. 30,* 5, 20–26.

LEVIS, P., PATEL, N., CULLER, D., AND SHENKER, S. 2004. Trickle: a self-regulating algorithm for code propagation and maintenance in wireless sensor networks. In *NSDI'04: Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation*, Berkeley, CA, USA, pp. 2–2. USENIX Association.

LIU, D. AND NING, P. 2003. Establishing pairwise keys in distributed sensor networks. In *CCS '03: Proceedings of the 10th ACM conference on Computer and communications security*, New York, NY, USA, pp. 52–61. ACM.

MALAN, D., WELSH, M., AND SMITH, M. 2004. A public-key infrastructure for key distribution in tinyos based on elliptic curve cryptography. In *SECON'04: IEEE International Conference on Sensor and Ad Hoc Communications and Network, Santa Clara, California, October 2004.*

MAY, R. M. AND LLOYD, A. L. 2001. Infection dynamics on scale-free networks. *Phys. Rev. E 64,* 6 (Nov), 066112.

MOORE, C. AND NEWMAN, M. E. J. 2000. Epidemics and percolation in small-world networks. *Phys Rev E 61,* 5, 5678–5682.

NEWMAN, M. E. J. 2002. The spread of epidemic disease on networks. *Physical Review Letters 66*, 016128.

NEWMAN, M. E. J., STROGATZ, S. H., AND WATTS, D. J. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E 64,* 026118.

PASTOR-SATORRAS, R. AND VESPIGNANI, A. 2001a. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E 63,* 6 (May), 066117.

PASTOR-SATORRAS, R. AND VESPIGNANI, A. 2001b. Epidemic spreading in scale-free networks. *Phys. Rev. Lett. 86,* 14 (Apr), 3200–3203.

PENROSE, M. D. 1999. On k-connectivity for a geometric random graph. *Random Struct. Algorithms 15,* 2, 145–164.

PIETRO, R. D., MEI, A., MANCINI, L. V., PANCONESI, A., AND RADHAKRISHNAN, J. 2006. Sensor networks that are provably resilient. In *Securecomm and Workshops '06: Proceedings of the 2nd IEEE International Conference on Security and Privacy for Emerging Areas in Communication Networks.*

STANIFORD, S., PAXSON, V., AND WEAVER, N. 2002. How to own the internet in your spare time. In *Proceedings of the 11th USENIX Security Symposium*, Berkeley, CA, USA, pp. 149–167. USENIX Association.

STAUFFER, D. 1985. *Introduction to percolation theory* (1 ed.). Taylor und Franci, London; Philadelphia.

WANG, L. 2004. Mnp: multihop network reprogramming service for sensor networks. In *SenSys '04: Proceedings of the 2nd international conference on Embedded networked sensor systems*, New York, NY, USA, pp. 285–286. ACM.

YAGAN, O. AND MAKOWSKI, A. M. 2008. On the random graph induced by a random key predistribution scheme under full visibility. In *ISIT'08: To appear at the IEEE International Symposium on Information Theory*, pp. 1–10.

YU, Z. AND GUAN, Y. 2005. A key pre-distribution scheme using deployment knowledge for wireless sensor networks. In *IPSN '05: Proceedings of the 4th international symposium on Information processing in sensor networks*, Piscataway, NJ, USA, pp. 35. IEEE Press.