electronics

MDPI

Article

Active Eavesdropping Detection Based on Large-Dimensional Random Matrix Theory for Massive MIMO-Enabled IoT

Li Xu [1,†], Jiaqi Chen [2,†], Ming Liu [1,*] and Xiaoyi Wang [3]

1 Beijing Key Lab of Transportation Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China; xuli16@bjtu.edu.cn
2 Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China; chenjq1016@hit.edu.cn
3 China Railway Fifth Survey and Design Institute Group Co. Ltd., Beijing 102600, China; wangxiaoyi@t5y.cn
* Correspondence: mingliu@bjtu.edu.cn
† These authors contributed equally to this work.

Received: 31 December 2018; Accepted: 28 January 2019; Published: 31 January 2019

Abstract: The increasing Internet-of-Things (IoT) applications will take a significant share of the services of the fifth generation mobile network (5G). However, IoT devices are vulnerable to security threats due to the limitation of their simple hardware and communication protocol. Massive multiple-input multiple-output (massive MIMO) is recognized as a promising technique to support massive connections of IoT devices, but it faces potential physical layer breaches. An active eavesdropper can compromises the communication security of massive MIMO systems by purposely contaminating the uplink pilots. According to the random matrix theory (RMT), the eigenvalue distribution of a large dimensional matrix composed of data samples converges to the limit spectrum distribution that can be characterized by matrix dimensions. With the assistance of RMT, we propose an active eavesdropping detection method in this paper. The theoretical limit spectrum distribution is exploited to determine the distribution range of the eigenvalues of a legitimate user signal. In addition the noise components are removed using the Marčenko–Pastur law of RMT. Hypothesis testing is then carried out to determine whether the spread range of eigenvalues is "normal" or not. Simulation results show that, compared with the classical Minimum Description Length (MDL)-based detection algorithm, the proposed method significantly improves active eavesdropping detection performance.

Keywords: Internet-of-Things; massive MIMO; active eavesdropper detection; random matrix theory

1. Introduction

With the prosperity of wireless networks and smart devices, more and more Internet-of-Things (IoT) applications have penetrated into various domains of industries, business, and daily lives of people. The fifth generation mobile network (5G) has taken the support of IoT applications as one of its major features [1]. However, IoT devices are vulnerable to the security threats from both network and physical layers due to the simplicity of the signal processing algorithm and encryption protocol running on them [2–4]. It is therefore, important to investigate the effective measures to counteract the increasing security threats.

Massive multiple-input multiple-output (massive MIMO) is seen as the key physical layer technique of 5G, and will bring unprecedented spectrum and energy efficiencies [1]. Recent works [5–10] have shown that the enormous degrees of freedom that massive MIMO brings can significantly increase the number of wireless connections, which makes massive MIMO a promising enabler for accommodating massive IoT devices. However, due to the openness of the wireless communication channel, massive MIMO technology also has security risks such as user signal

eavesdropping and information leakage. Existing systems usually use encryption algorithms in high-level protocols to ensure user information security. With the continuous development of computing capacity, the communication security solely relying on the high-level encryption is facing more and more challenges [11,12].

Physical layer security technology incorporates advanced signal processing methods such as beamforming and artificial noise at the physical layer and can reduce the possibility of legitimate users' information acquired by malicious users [13–15]. The strong spatial beamforming capability of massive MIMO leads to good resistance to passive eavesdropping. However, massive MIMO systems rely on deterministic pilot symbol sequences in the uplink training for channel information acquisition. Once the deterministic sequence is acquired by a malicious user, it can actively transmit this sequence at the same time as the legitimate user. This procedure will mislead the channel estimation process of the base station, and redirect the downlink signal beam to the malicious user's position. This type of attack is often referred to as active eavesdropping [16].

Zhou et al. pointed out the security risks caused by active eavesdropping to wireless communication systems with time division duplex (TDD) transmission [11]. Some studies were conducted to deal with these problems [12,16–24], including the random sequence based detection algorithms [16,20,25], channel statistics based detection algorithms [21,22], signal power based detection algorithm [18,23] and signal subspace-based detection algorithm [24]. Among them, the signal subspace-based detection method is an effective way to detect active eavesdropping. The general idea is to determine whether the number of detected system signal sources is equal to the number of legitimate users. If yes, no active eavesdropping is undertaking in the system. However, if the number of detected signal sources is greater than the number of legitimate users, there is high possibility that the system is under the attack of active eavesdropping.

The existing eavesdropping detection algorithms leveraging signal subspace are based on information theory criteria for eavesdropping detection, such as minimum description length (MDL) [26,27] and Akaike information criterion (AIC) [28]. The algorithm introduced by Wax and Kailath [29] is used for adaptive detection of eavesdropping users in a parameter changing environment. However, the penalty terms in detection criteria affect the accuracy of the detection. For example, the MDL algorithm has underestimation problems, while the AIC algorithm has over estimation problems, caused by their penalty terms [24].

More recently some methods were proposed to employ random sequences for the active eavesdropping purpose [16,25]. The rationale behind these methods is that the random sequences are not deterministic and cannot be acquired through historical observation, nor predicted through time series techniques by malicious users. Hence, if the statistic property of the random sequence is properly characterized, the active eavesdropping can be perceived by detecting the anomaly in the statistics of received random sequence. These methods are particularly suitable for the IoT applications because their simplicity can meet the strict complexity and energy constraint of IoT devices [2,3]. However, they were all designed for the classical single user antenna assumption. With the development of massive MIMO technology, the transition from single user antenna to multiple user antennas is an inevitable trend. Therefore, we need to study the active eavesdropping detection mechanism for multi-antenna scenarios.

With the general assumption that mobile users employ multiple antennas, this paper proposes to combine the signal subspace based and random sequence base active eavesdropping detection ideas. Random sequences are transmitted to create the random user features that cannot be forged by eavesdropper. Consequently the limit distribution of the eigenvalues derived from large-dimensional random matrix theory (RMT) [30,31] is used to calculate the boundaries of the sample distribution so that the region of the "normal" samples falling into can be properly characterized. Then the decision statistics is designed accordingly.

The rest of the paper is organized as follows. The active eavesdropping problem is described in Section 2. Section 3 proposes an active eavesdropping detection algorithm based on RMT. Simulation results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Active Eavesdropping in Massive MIMO Systems

This paper considers a scenario where base station *Alice* has a large amount of (M) antennas and each legitimate user *Bob* equips multiple (K) antennas. The active eavesdropping is shown in Figure 1. The legitimate user attempts to setup the wireless link with the base station. To this end Bob sends a series of orthogonal training sequences through its antennas to Alice so that Alice is able to estimate the channel coefficients based on these known sequences. The eavesdropper *Eve* tries to affect the channel estimation process of Alice via sending the same training sequences as Bob. Without loss of generality we assume that Eve has the same number of antennas as Alice. When the number of antennas of the eavesdropping user is greater than or equal to the number of antennas of the legitimate user, the information of the legitimate user can be effectively demodulated. However, the more the number of the eavesdropper's antennas, the easier it is to be detected. Therefore, from the perspective of eavesdropper detection, we consider the most difficult situation, that is, the eavesdropper's antennas is equal to the legitimate user's antennas.

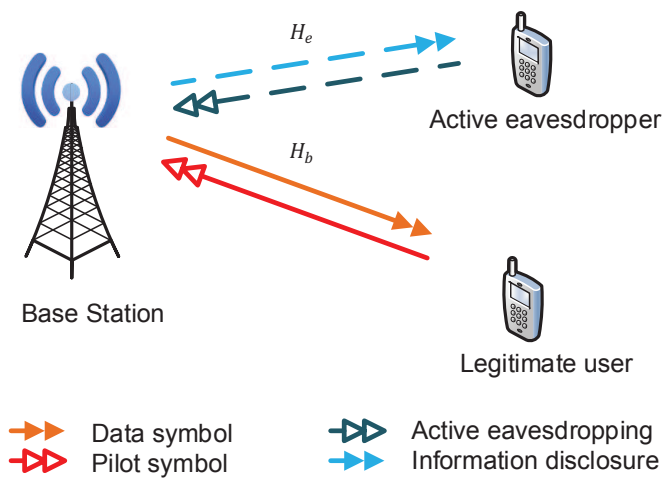


Figure 1. An illustration of active eavesdropping in massive MIMO-based systems.

Following the random sequence based active eavesdropping detection method, a series of random sequences are transmitted after the training sequences. The random sequences received by Alice with and without the contamination from Eve are expressed as:

$$y_{b_j} = \sqrt{p_b^t}\sqrt{\alpha}\sum_{i=1}^{K} h_{b_{ji}} x_{b_i} + n_{b_j}, \tag{1}$$

$$y_{e_j} = \sqrt{p_b^t}\sqrt{\alpha}\sum_{i=1}^{K} h_{b_{ji}} x_{b_i} + \sqrt{p_e^t}\sqrt{\beta}\sum_{i=1}^{K} h_{e_{ji}} x_{e_i} + n_{e_j}, \tag{2}$$

where x_{b_i} and x_{e_i} represent the random sequence transmitted from Bob and Eve, respectively, by the ith transmit antenna; y_{b_j} and y_{e_j} represent the random sequence received by the jth receive antenna with and without impact from Eve, respectively; $h_{b_{ji}}$ and $h_{e_{ji}}$ are the channel fading coefficient betweens the ith transmit antenna of Bob or Eve and the jth receive antenna at base station, respectively. Real values α and β represent the associated path-loss attenuation factors. p_b^t and p_e^t are the uplink transmission power of Bob and Eve, respectively.

Rewrite the above Equations (1) and (2) in matrix form:

$$Y_b = \sqrt{p_b^t}\sqrt{\alpha}\widetilde{H_b}X_b + N_b = \sqrt{p_b}\widetilde{H_b}X_b + N_b, \tag{3}$$

$$Y_e = \sqrt{p_b^t}\sqrt{\alpha}\widetilde{H_b}X_b + \sqrt{p_e^t}\sqrt{\beta}\widetilde{H_e}X_e + N_e = \sqrt{p_b}\widetilde{H_b}X_b + \sqrt{p_e}\widetilde{H_e}X_e + N_e, \tag{4}$$

where $X_b \in \mathbb{C}^{K\times T}$ and $X_e \in \mathbb{C}^{K\times T}$ are the random sequences transmitted from Bob and Eve, with the sequence length of T; $Y_b \in \mathbb{C}^{M\times T}$ and $Y_e \in \mathbb{C}^{M\times T}$ are the sequences received by Alice when there exists and does not exist contamination from Eve, respectively; $\widetilde{H_b} \sim \mathcal{CN}(0, I_M)$ and $\widetilde{H_e} \sim \mathcal{CN}(0, I_M)$ represent small-scale fading. $p_b = \alpha p_b^t$ and $p_e = \beta p_e^t$ are the received signal power of legitimate user and eavesdropper, respectively. $N_b \in \mathbb{C}^{M\times T}$ and $N_e \in \mathbb{C}^{M\times T}$ represent the Gaussian noise in the two cases. Their elements are independent and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and variance of σ^2. We assume that the base station can accurately estimate the noise level through long-term observation [32,33]. Since X_b and X_e are randomly generated in an independent manner, they will expand the K-dimensional subspaces, respectively.

Denote $H_b = \sqrt{\alpha}\widetilde{H_b} \in \mathbb{C}^{M\times K}$, $H_e = \sqrt{\beta}\widetilde{H_e} \in \mathbb{C}^{M\times K}$, $X_b'(t) = \sqrt{p_b^t}X_b(t)$ and $X_e'(t) = \sqrt{p_e^t}X_e(t)$. The active eavesdropping detection problem can be characterized by the following hypothesis test:

$$\mathcal{H}_0 : Y(t) = H_b(t)X_b'(t) + N_b(t), \tag{5}$$

$$\mathcal{H}_1 : Y(t) = H_b(t)X_b'(t) + H_e(t)X_e'(t) + N_e(t). \tag{6}$$

Wherein, it is assumed that the channel is quasi-static, i.e., channel fading coefficients being constant for a period of at least T consecutive samples.

3. Active Eavesdropping Detection Based on Large-Dimensional Random Matrix Theory

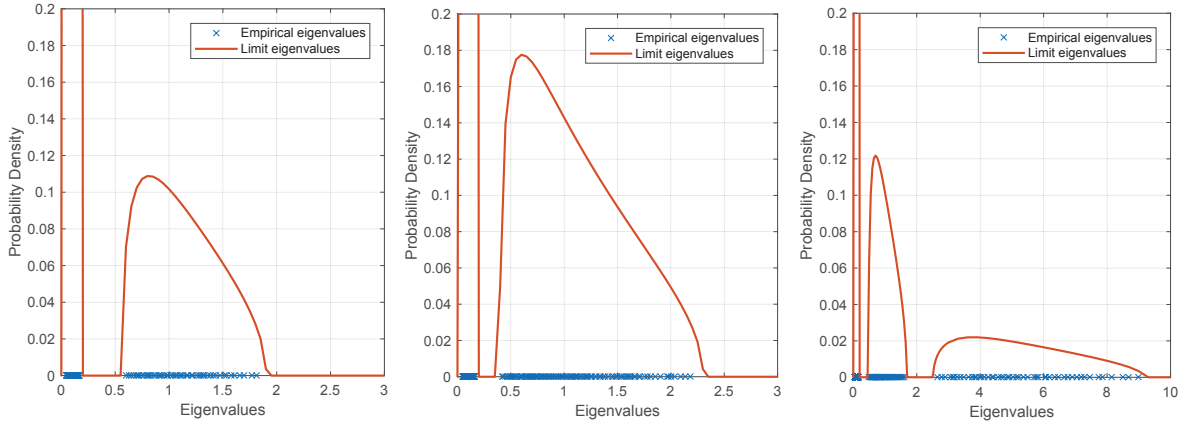
3.1. Rational Behind the Proposal

According to the idea of signal subspace, if there exist several independent components in the received signal, the spectral distribution of the eigenvalues is near the position characterized by its power, when the number of antennas at the base station and user is large. The empirical spectral distribution of each component converges to its limit distribution. The expression of the eigenvalue distribution can be derived from the large-dimensional RMT. This provides an idea of using the limit distribution as a theoretical guideline to check whether the empirical eigenvalue distribution in practice is reasonable.

Examples of the empirical eigenvalue distribution of B_N are presented in Figure 2. We can observe that the eigenvalues fall into a larger support when there is one eavesdropper compared with the case when there is not, even though the signal powers of user and eavesdropper are the same. Moreover, the eigenvalues fall into multiple disjoint supports when the powers of user and eavesdropper are significantly different. This provides an intuitive evidence that the distribution of the eigenvalues can be leveraged to determine the existence of active eavesdropping.

The large-dimensional RMT suggests that, with the increase of the number of base stations and user antennas, the empirical spectral distribution of the received signal of the base station will gradually converge to its limit spectral distribution. That is, the distribution of empirical spectrum can be accurately estimated when the matrix dimension is large, i.e., M, K, $T \to \infty$. Note that in reality, the quantities of M, K and T just need to be "sufficiently" large, say on the level of several tens to hundreds, to validate RMT. This assumption can be approaches in the massive MIMO scenarios. Therefore, the theoretical distribution of the limit spectrum can be used to judge whether the observed empirical spectral distribution is "normal". This constitutes the basis of using the large-dimensional RMT to detect active eavesdropping in massive MIMO systems.

In the following sections, we first characterize the eigenvalue distribution properties of received massive MIMO data sequence before proposing the detection mechanism.



(**a**) (**b**) (**c**)

Figure 2. (**a**) Empirical distribution of eigenvalues of massive MIMO when there is no eavesdropper $p_b = 1$, $\frac{T}{M} = 10$, $\frac{M}{K} = 10$, $\sigma^2 = 0.1$. (**b**) There is one active eavesdropper with equal received power, i.e., $p_b = p_e = 1$, and $\frac{T}{M} = 10$, $\frac{M}{K} = 10$, $\sigma^2 = 0.1$. (**c**) There is one active eavesdropper with different received power, i.e., $p_b = 1$, and $p_e = 5$, $\frac{T}{M} = 10$, $\frac{M}{K} = 10$, $\sigma^2 = 0.1$.

3.2. Eigenvalue Distribution of $(\frac{1}{T}YY^H)$

Recall the random sequences received by base station $Y \in \mathbb{C}^{M\times T}$. We provide an analysis of the eigenvalue distribution of the sample covariance matrix $B_N = \frac{1}{T}YY^H$ from the large-dimensional RMT in the sequel.

Let $F(t)$ be the population spectral distribution function. The empirical spectral distribution (e.s.d.) of B_N, noted by $F^{(B_N)}$, converges to the deterministic, limit spectral distribution (l.s.d.) F of B_N. We have following expression [30]:

$$z(m) = -\frac{1}{m} + \frac{1}{c}\int \frac{t}{1+tm} dF(t), \tag{7}$$

where m is the Stieltjes transform of F with $t \in \mathbb{R}$, $z \in \mathbb{C}^+$. When M, K and T grow at the same rate, i.e., $M, K, T \to \infty$ with constant $c = \frac{T}{M}$. The Stieltjes transform has one-to-one correspondence to the distribution of the eigenvalues of the matrix. Mathematically it is easier to use and can build the relationship to the linear spectral statistics and other related properties.

The sample covariance matrix B_N has an overall covariance matrix of uncertainty, but its empirical spectral distribution has an almost certain limit. Accordingly, the problem is transformed into a limit distribution that solves the empirical spectral distribution.

For the sake of conciseness, we first derive the signal expression when there is one legitimate user. Following the system model considered in previous section, the received signal is:

$$Y = HP^{1/2}X + \sigma W, \tag{8}$$

where σW is the noise components. $H \in \mathbb{C}^{M\times K}$, $P \in \mathbb{C}^{K\times K}$, $X \in \mathbb{C}^{K\times T}$. Rewrite the above expression:

$$Y = \begin{pmatrix} HP^{1/2} & \sigma I_N \end{pmatrix} \begin{pmatrix} X \\ W \end{pmatrix}. \tag{9}$$

Add Y to zero vector 0 to get a larger matrix $\underline{Y} \in \mathbb{C}^{(M+K)\times T}$:

$$\underline{Y} = \begin{pmatrix} HP^{\frac{1}{2}} & \sigma I_N \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ W \end{pmatrix}, \tag{10}$$

where K refers to the sum of the number of legitimate user's antennas. We can find that $\frac{1}{T}(\underline{YY^H})$ is a sample covariance matrix, for which the population covariance matrix is:

$$\begin{pmatrix} HPH^H + \sigma^2 I_N & 0 \\ 0 & 0 \end{pmatrix}. \tag{11}$$

It is nondeterministic and the random matrix $\binom{X}{W}$ has independent (unnecessarily identically distributed) entries with zero mean and variance of one. When $K, M \to \infty$, H is a term with independent finite fourth-order moments. P is a diagonal matrix whose entries are the source powers with multiplicities the number of transmit antennas of each user.

Let Q_{HPH^H} denote the distribution function of the eigenvalues of HPH^H. It is shown that $m_\infty(z)$ is the Stieltjes transform of a distribution function Q_∞ which is the limit of Q_{HPH^H}:

$$m_\infty(z) = \int \frac{1}{\lambda - z} dQ_\infty(\lambda), \tag{12}$$

where z is the inverse of the Stieltjes transform, and is calculated as:

$$z = -\frac{1}{m_\infty} + \frac{1}{c} \int \frac{t}{1 + t m_\infty} dF_\infty(t). \tag{13}$$

As introduced in (7), $F(t)$ is the population spectral distribution function, and it converges in distribution to a distribution $F_\infty(t)$. For finite T and M, m satisfies:

$$z(m) = -\frac{1}{m} + \frac{1}{c} \int \frac{t}{1 + tm} dF(t) + \zeta, \tag{14}$$

where m is the Stieltjes transform of Q_{HPH^H} and $m \to m_\infty$. The "error term" $\zeta \to 0$ as $M \to \infty$. Hence (14) becomes (13). In Section 3.3 we will derive an estimate of ζ, and show that by setting $\zeta = 0$ in (14) and solving m, we will obtain a finer asymptotics of the eigenvalues of HPH^H.

The almost sure convergence of the e.s.d. of HPH^H ensures the almost sure convergence of the e.s.d. of the matrix in (11). The distribution range of eigenvalues of HPH^H can be obtained by z in (14). Then, we consider a special case, when $P = \sigma^2 I_N$, we can get the boundary range of the noise eigenvalues. Thus, the distribution of the eigenvalues of $HPH^H + \sigma^2 I_N$ can be obtained by integrating the ranges of HPH^H and noise terms. In the following sections the detailed derivation of the eigenvalues' distribution range will be presented.

3.3. Eigenvalues' Boundary of (HPH^H)

We have an empirical spectral distribution of HPH^H that weakly and almost surely converges to the limit distribution G. Therefore, we can first obtain the limit distribution of HPH^H.

With some trivial manipulation we can generalized expression (7) to the case of L users. Denote the Stieltjes transform of G as m_G. Its inverse transform is $x_G(m_G) \in \mathbb{C}^+$ [17]:

$$x_G(m_G) = -\frac{1}{m_G} + \sum_{r=1}^{L} \frac{p_r}{c_r(1 + p_r m_G)}, \tag{15}$$

where $n_1, n_2, \cdots, n_L$ are the number of antennas of the first, the second, $\cdots$, and the Lth legitimate user, $c_r = \frac{M}{n_r} > 0, K = n_1 + n_2 + \cdots + n_L$.

In Equation (15), we give the inverse of the Stieltjes transform when there are multiple legitimate users. Assuming that there is only one legitimate user in the system, (15) is turned to:

$$x_G(m_G) = -\frac{1}{m_G} + \frac{p_b}{c_1(1 + p_b m_G)}. \tag{16}$$

The discretized minimal and maximal values of the eigenvalues according to the limit distribution are:

$$x_G{}^- = -\frac{1}{m_G{}^-} + \frac{p_b}{c_1(1+p_b m_G{}^-)}, \tag{17}$$

$$x_G{}^+ = -\frac{1}{m_G{}^+} + \frac{p_b}{c_1(1+p_b m_G{}^+)}, \tag{18}$$

where $m_G{}^-$ and $m_G{}^+$ are the two real-valued solutions of the equation:

$$\frac{(p_b m_G)^2}{c_1(1+p_b m_G)^2} = 1, \tag{19}$$

and satisfies $m_G{}^- < m_G{}^+$.

Now we briefly reveal the derivation of Equation (19). The boundary of the eigenvalues is at the extreme point of the integral curve $x_G(m_G)$ [17], from which we have:

$$x'_G(m_G) = \frac{1}{m_G{}^2}\left[1 - \sum_{r=1}^{L} \frac{p_r{}^2}{c_r} \frac{1}{(p_r + \frac{1}{m_G})^2}\right]. \tag{20}$$

Therefore, let $x'_G(m_G) = 0$, we can get:

$$\sum_{r=1}^{L} \frac{p_r{}^2}{c_r} \frac{1}{(p_r + \frac{1}{m_G})^2} = 1. \tag{21}$$

Rearranging the above equation, we can get the Equation (19).

From these expressions we can reach an important observation: if we want to increase the estimator's detection sensitivity, that is, to be able to separate two signal sources with similar received power, we need to increase the number of base station antennas (yielding a larger c_1).

However, the left and right boundaries of the eigenvalues obtained in (17) and (18) have some deviations. As shown in Figure 3, eigenvalues fall in the areas outside the boundary $[x_G{}^-, x_G{}^+]$ (i.e., outside the areas bounded by blue circles).

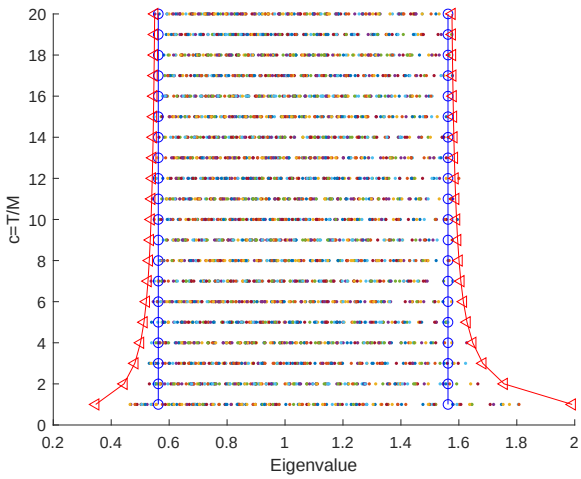


Figure 3. Empirical eigenvalue distribution without the effect of noise, when $p_b = 1$ and $\frac{M}{K} = 10$. The blue circles represent the boundaries of the eigenvalues obtained from (17) and (18). The red triangle represent the boundaries from (25) and (26) that incorporate the correction item ε.

As shown in Figure 2, when the matrix dimension is large and the powers of the two signal sources satisfy some requirements, the supports of the eigenvalues associated to the two sources are separable. According to the separability of eigenvalues' supports derived in [31] in which some corrections were proposed to improve the boundaries where there is no eigenvalue. We have the following Lemma.

Lemma 1. *Denote the interval between two supports of eigenvalues as $[a, b]$, where there is no eigenvalue.*

$$P(\lambda_{i_M} > b + \varepsilon',\ \lambda_{i_{M+1}} < a - \varepsilon') = 1, \tag{22}$$

where λ_{i_M} is the eigenvalue of the i_Mth eigenvalue cluster, and $\lambda_{i_{M+1}}$ is the eigenvalue of the i_{M+1}th eigenvalue cluster, and the eigenvalues are arranged in a descending order, i.e., $\lambda_{i_M} > \lambda_{i_{M+1}}$. ε' is calculated as:

$$\varepsilon' = \frac{1}{\widehat{a}|m_a|}, \tag{23}$$

and

$$\widehat{a} = a + c\left|\int \frac{t}{1 + tm_a} dF(t)\right|. \tag{24}$$

According to the above Lemma, we propose to modify (17) and (18) with a similar correction term ε:

$$x_1 = -\frac{1}{m_G{}^- - \varepsilon} + \frac{p_b}{c_1(1 + p_b(m_G{}^- - \varepsilon))} - \varepsilon, \tag{25}$$

$$x_2 = -\frac{1}{m_G{}^+ + \varepsilon} + \frac{p_b}{c_1(1 + p_b(m_G{}^+ + \varepsilon))} + \varepsilon. \tag{26}$$

The correction term ε is an infinitesimal, and $\varepsilon > 0$. In this case, ε in fact represents the ζ in Equation (14). It can be calculated as:

$$\varepsilon = \frac{1}{\widehat{x_G{}^+}|m_G{}^+|}, \tag{27}$$

where

$$\widehat{x_G{}^+} = x_G{}^+ + c\left|\frac{p_b}{1 + p_b m_G{}^+}\right|. \tag{28}$$

With these new expressions in (25) and (26), we obtain new support of eigenvalues:

$$\left(x_G{}^-(m_G{}^- - \varepsilon) - \varepsilon,\ \ x_G{}^+(m_G{}^+ + \varepsilon) + \varepsilon\right). \tag{29}$$

Observing Figure 3, we find that the corrected eigenvalue boundaries are more accurate compared to the expression given in (17) and (18).

3.4. Eigenvalues' Boundary of $(HPH^H + \sigma^2 I_M)$

The left and right boundaries x_1 and x_2 obtained from (25) and (26) are only based on the empirical spectral distribution of HPH^H. They characterize the range of the received signal eigenvalues in the absence of noise. With the effect of additive noise, the range of eigenvalues has a significant shift to the right, and the boundaries x_1 and x_2 no longer conform to the eigenvalue distribution, as shown in Figure 4. With this observation, we continue to improve the expression of the eigenvalues' boundaries by incorporating the noise effect to (25) and (26).

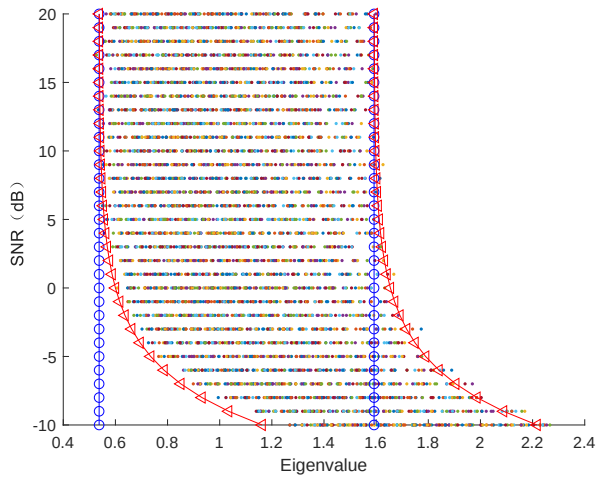


Figure 4. Empirical eigenvalue distribution with the effect of noise at different SNR levels when $p_b = 1$, $\frac{T}{M} = 10$, $\frac{M}{K} = 10$. The blue circles represent the boundaries of the eigenvalues obtained from (25) and (26). The red triangle represent the boundaries from (30) and (31) that integrate the effect of noise term.

In the massive MIMO scenarios considered in this work, the length of the support of HPH^H is approximately equal to the length of the support of $HPH^H + \sigma^2 I_M$. The main difference between them is that the latter has a significant offset from the former due to the influence of noise. Unfortunately it is difficult to analytically characterize the influence of the noise. We propose the following modifications to the eigenvalues' boundary expressions:

$$x_l = -\frac{1}{m_G{}^- - \varepsilon} + \frac{p_b}{c_1(1 + p_b(m_G{}^- - \varepsilon))} - \varepsilon + x_{nG}, \tag{30}$$

$$x_r = -\frac{1}{m_G{}^+ + \varepsilon} + \frac{p_b}{c_1(1 + p_b(m_G{}^+ + \varepsilon))} + \varepsilon + x_{nG}, \tag{31}$$

where the noise effect is counted in the term:

$$x_{nG}{}^- = -\frac{1}{m_{nG}{}^-} + \frac{\sigma^2}{c_1(1 + \sigma^2 m_{nG}{}^-)}, \tag{32}$$

$$x_{nG}{}^+ = -\frac{1}{m_{nG}{}^+} + \frac{\sigma^2}{c_1(1 + \sigma^2 m_{nG}{}^+)}, \tag{33}$$

$$x_{nG} = x_{nG}{}^+ - x_{nG}{}^- \tag{34}$$

in which m_{nG} is obtained by:

$$\frac{(\sigma^2 m_{nG})^2}{c_1(1 + \sigma^2 m_{nG})^2} = 1. \tag{35}$$

We plot the new boundaries in Figure 4. It is clear that the boundaries proposed in (30) and (31) fit better the real distribution of the eigenvalues compared to the expressions in (25) and (26).

Upon this basis, we propose a test criterion for detecting active eavesdropping:

$$\gamma = x_r - x_l. \tag{36}$$

This value in fact characterizes the maximal range that the eigenvalues distribute when there is only legitimate user. If the actual spread of the eigenvalues is greater than this range, it suggests that there exists some anomaly in the received signal, indicating the active eavesdropping phenomenon.

3.5. Noise Elimination Based on Marčenko–Pastur Law

In this section, we propose to use the limit distribution of the noise covariance matrix to approximate its empirical distribution. More precisely we use the Marčenko–Pastur distribution from large-dimensional RMT (commonly referred to as M-P law) to help eliminating the noise components.

The noise component N in the received signal of the base station satisfies complex Gaussian distribution, i.e., $N_{ij} \sim \mathcal{CN}(0, \sigma^2)$. Then, according to M-P law, the left and right boundaries of its limit spectral distribution can be written as:

$$n_l = \sigma^2 (1 - \sqrt{\frac{M}{T}})^2, \tag{37}$$

$$n_r = \sigma^2 (1 + \sqrt{\frac{M}{T}})^2. \tag{38}$$

With this knowledge, we can eliminate the eigenvalues of the received signal that fall into the range $[n_l, n_r]$.

The use of the M-P law to eliminate noise components will bring the following advantages. As long as the noise power and legitimate user's power are relative different, the empirical spectral distribution of noise and legitimate user's signal are clearly separated. This method does not affect the eigenvalues brought by the active eavesdropping signal components, and thereby does not affect the consequent active eavesdropping detection. In contrast, other traditional noise elimination method that removes a certain number of smallest eigenvalues may cause potential risk of miss detection of active eavesdropping [29]. This is because the eigenvalues associated with the active eavesdropper may be removed in this process, which will jeopardize the following eavesdropping detection process.

3.6. Proposed Active Eavesdropping Detection Algorithm

Based on the previous analysis and discussion, we propose the following algorithm to detect active eavesdropping.

Step 1: **Preliminary**. The base station calculates a theoretical decision threshold of the test statistic $\gamma = x_r - x_l$ based on the system parameters e.g., base station's antenna number M, user's antenna number K, signal sample number T, as well as the historical observation of noise level σ^2 and user signal power p_b.

Step 2: **Eigenvalue computation**. The base station samples the received random sequence and obtains a sample matrix Y. Then it calculates the sample covariance matrix $B_N = YY^H/T$ and performs the eigenvalue decomposition of B_N. This yields M eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_M > 0$.

Step 3: **Noise elimination**. The base station uses the M-P law to compute the noise range $[n_l, n_r]$. All the eigenvalues falling into this range are eliminated.

Step 4: **Test statistic calculation**. The difference between the maximal and minimal values of the remaining eigenvalues is calculated, i.e., $\bar{\gamma} = \lambda_1 - \lambda_{K'}$ where $\lambda_{K'}$ is the smallest remaining eigenvalue.

Step 5: **Hypothesis test**. The test statistic $\bar{\gamma}$ obtained in Step 4 is compared to the theoretical range of the eigenvalue spread γ. If $\bar{\gamma} < \gamma$, the null hypothesis $\mathcal{H}_0$ is accepted. This means there is no active eavesdropping components in the received signal. In the contrary, if $\bar{\gamma} > \gamma$, the alternative hypothesis $\mathcal{H}_1$ is accepted. It suggests that it is highly probable that the system is under active eavesdropping.

By periodically performing the previous steps, the system can detect the existence of active eavesdropping. We can get the Algorithm 1 shows the process of active eavesdropping detection.

Algorithm 1 Active Eavesdropping Detection Algorithm

Input: Base station received signal Y as shown in (3) and (4).

Output: Active eavesdropping detection result.

1: Calculate test statistic $\gamma = x_r - x_l$ as shown in (30) and (31).
2: Solve the eigenvalue of B_N $\lambda_1 > \lambda_2 > \ldots > \lambda_M > 0$.
3: Eliminate the eigenvalues caused by noise in $[n_l, n_r]$ as shown in (33).
4: Calculate the actual value of the test statistic.
5: Perform a hypothesis test based on the test statistic.

4. Performance Evaluation

In this section, we evaluate the performance of the proposed algorithm with various parameters M, p_b and T based on the analytical expressions derived in Sections III and via Monte Carlo simulations. The detection probability indicates the probability that the eavesdropping user is detected, and the false alarm probability indicates that the probability of the eavesdropper is erroneously detected when there is no eavesdropper. All relevant system parameters are provided in the captions of figures. Rayleigh flat-fading channel is adopted in the simulations.

In Figure 5, we show the detection probability with respect to signal-to-noise ratio (SNR) levels for different base station antennas M. The detection probability increases to one and the alarm probability tends to zero when SNR increases. When noise power is equal to the power of legitimate user signal, the overlap of the empirical spectral distribution of the signal and noise components is more serious. As SNR increases, the empirical spectral distribution of user signal and noise components is gradually separated, and the detection performance is significantly improved. Moreover, the proposed algorithm obviously outperforms the classical MDL-based detection method [24].

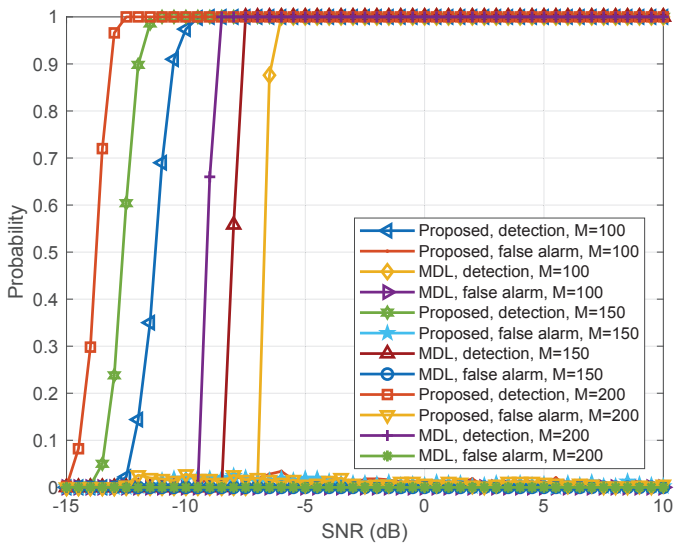


Figure 5. Effect of the number of base station antennas on detection performance when $p_b = 1$, $p_e = 1$, $\frac{T}{M} = 10$ and $K = 16$.

Figure 6 shows the detection probability with respect to SNR with different numbers of data samples. The experimental results show that the detection algorithm is improved when employing more data samples. Moreover, traditional detection methods [24] require a particularly large number of samples to approximate the sample covariance matrix. However, the proposed scheme only needs to observe the spectrum of the large-scale sample covariance matrix. Even when the number of samples is not very large, we can still obtain a relatively good approximate of the spectrum of sample covariance matrix. From the figure, it is clearly that when the value of c is relatively low (e.g., 0.8), the new left and right boundaries can still characterize the actual eigenvalue distribution. In particular, when the

number of samples and the number of signal sources are in the same order of magnitude, the proposed scheme can effectively detect the active eavesdropping.

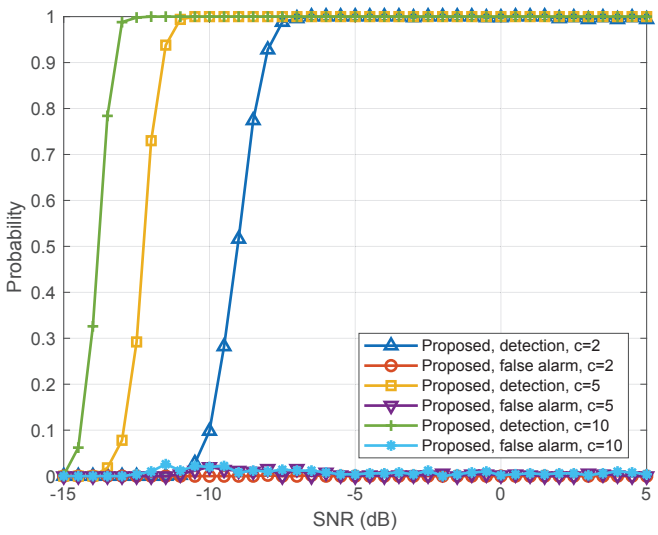


Figure 6. Effect of number of samples on algorithm performance when $p_b = 1$, $p_e = 1$, $M = 256$ and $K = 16$.

Figure 7 illustrates the detection probability with respect to the relative power ratio of the active eavesdropper versus legitimate user (p_e/p_b). As expected, with the increase of p_e, detection probability is significantly improved. We should note here that the portion of base station signal redirected to eavesdropper linearly related to eavesdropper's transmit power. Therefore, there is no interest for eavesdropper to unlimitedly reduce its transmit power to purposely avoid from being detected. So, the fact that the proposed scheme can work in very low SNR regime (e.g., -10 dB) and low p_e/p_b ratio suggests that the eavesdropper must have highly superiority over legitimate user (such as having much higher sensitivity or being much closer to the base station) to make the eavesdropping possible.

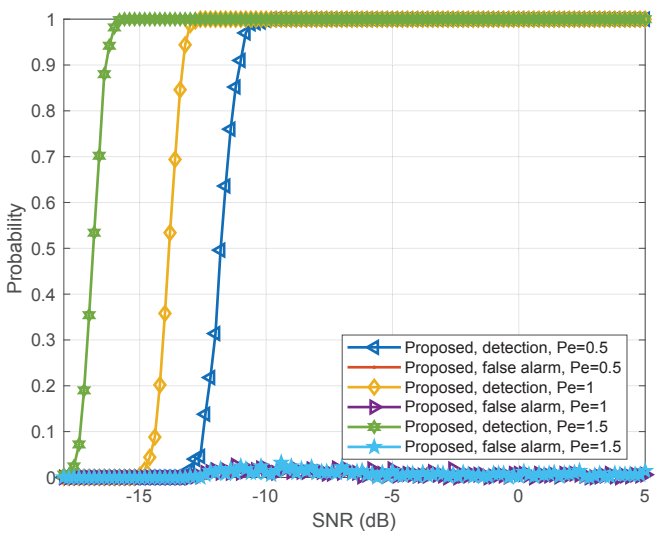


Figure 7. Effect of eavesdropper's signal andlegitimate user's signal power on detection performance when $p_b = 1$, $\frac{T}{M} = 10$, $M = 256$ and $K = 16$.

In Figure 8, we conduct the simulation to illustrate the performance of the proposed algorithm as a function of eavesdropper's power p_e. We set the legitimate user's power $p_b = 1$ and $p_b = 3$, respectively, and the probability of false alarm as $P_{fa} = 0.1$. From Figure 8, we can conclude that

higher eavesdropper power leads to higher probability of correct attack detection. More importantly, even with weak eavesdropper signal, for example $p_e = 0.04$, the performance of our proposed scheme is still high accuracy. This suggests that the proposed scheme can detect the eavesdropping even when the power of eavesdropper is much lower that the legitimate user. We also compare the energy ratio (ER) based algorithm [23] and MDL based algorithm [24]. Obviously, the proposed algorithm significantly outperforms ER and MDL based ones.

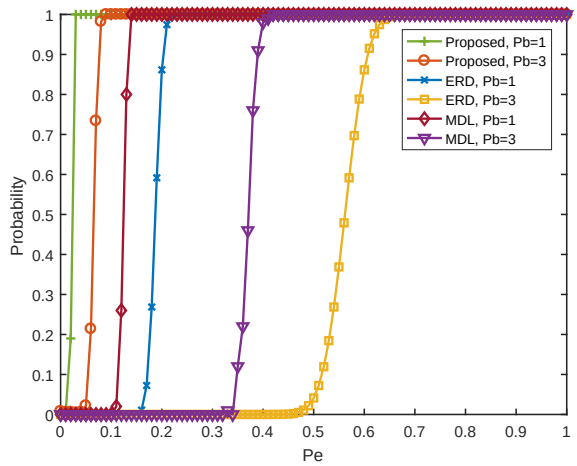


Figure 8. Effect of eavesdropper's signal power p_e on detection performance when $\frac{T}{M} = 10$, $M = 256$, $K = 16$ and $SNR = -2$ dB.

In Figure 9, the impact of legitimate user's and eavesdropper's antenna numbers on the detection performance is evaluated. From the figure, as the number of eavesdropper's antennas increases, the performance of the proposed algorithm is improved. Recall that only when the number of eavesdropper's antennas is greater than or equal to the number of legitimate user's antennas, the eavesdropper can effectively eavesdrop information. Therefore, the results suggest that the proposed scheme can effectively detect active eavesdropping with the reasonable antenna number settings.

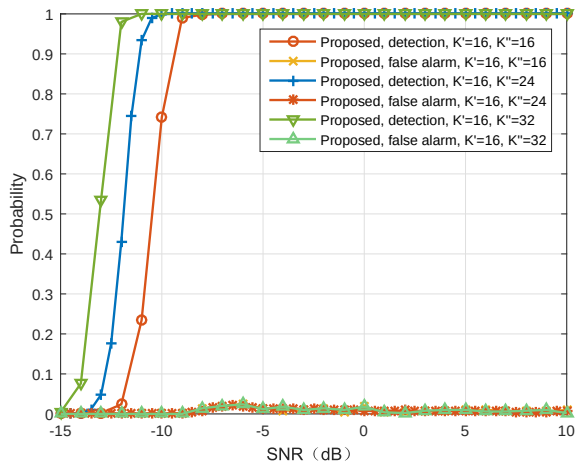


Figure 9. Effect of eavesdropper's signal and legitimate user's antenna number on detection performance when $p_b = 1$, $\frac{T}{M} = 10$, $M = 128$ and $K = 16$.The parameter K' represents the number of antennas of the legitimate user, and the parameter K'' represents the number of antennas of the eavesdropper.

5. Conclusions

In this paper, we proposed an active eavesdropping detection algorithm based on large-dimensional RMT for massive MIMO enabled systems. The algorithm uses the limit spectral distribution of eigenvalues as a theoretical criterion to determine whether the distribution of the eigenvalues of the received signal is normal, that is, whether it contains an active eavesdropping signal component. Compared with existing detection algorithms, the proposed scheme is able to achieve more reliable and accurate detection performance in low SNR scenarios and the number of samples needed can be smaller than the number of antennas.

Author Contributions: Conceptualization, M.L., X.W. and J.C.; methodology, L.X., M.L. and J.C.; validation, L.X., M.L., J.C. and X.W.; formal analysis, L.X. and J.C..

Funding: L. Xu and M. Liu's work is funded by National Natural Science Foundation of China (No. 61501022). J. Chen's work is supported by the National Natural Science Foundation of China (No. 11501147, 91646106).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike Information Criterion
e.s.d.	Empirical Spectral Distribution
IoT	Internet of Things
l.s.d.	Limit Spectral Distribution
MDL	Minimum Description Length
MIMO	Multiple-Input Multiple-Output
M-P Law	Marčenko–Pastur Law
RMT	Random Matrix Theory
SNR	Signal-to-Noise Ratio
TDD	Time Division Duplex

References

1. Shafi, M.; Molisch, A.F.; Smith, P.J.; Haustein, T.; Zhu, P.; Silva, P.D.; Tufvesson, F.; Benjebbour, A.; Wunder, G. 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment and Practice. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1201–1221. [CrossRef]
2. Mukherjee, A. Physical-Layer Security in the Internet of Things: Sensing and Communication Confidentiality Under Resource Constraints. *Proc. IEEE* **2015**, *103*, 1747–1761. [CrossRef]
3. Burg, A.; Chattopadhyay, A.; Lam, K. Wireless Communication and Security Issues for Cyber–Physical Systems and the Internet-of-Things. *Proc. IEEE* **2018**, *106*, 38–60. [CrossRef]
4. Hamamreh, J.M.; Furqan, H.M.; Arslan, H. Classifications and Applications of Physical Layer Security Techniques for Confidentiality: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2018**. [CrossRef]
5. Deng, R.; Zhou, S.; Niu, Z. Scalable Non-Orthogonal Pilot Design for Massive MIMO Systems with Massive Connectivity. In Proceedings of the IEEE Global Communications Conference (Globecom) Workshops, Washington, DC, USA, 4–8 December 2017; pp. 1–6.
6. Lee, B.M.; Yang, H. Massive MIMO for Industrial Internet of Things in Cyber-Physical Systems. *IEEE Trans. Ind. Inform.* **2018**, *14*, 2641–2652. [CrossRef]
7. Deng, R.; Jiang, Z.; Zhou, S.; Niu, Z. How Often Should CSI Be Updated for Massive MIMO Systems with Massive Connectivity? In Proceedings of the IEEE Global Communications Conference (Globecom), Singapore, 4–8 December 2017; pp. 1–6.
8. Liu, L.; Yu, W. Massive Connectivity with Massive MIMO—Part I: Device Activity Detection and Channel Estimation. *IEEE Trans. Signal Process.* **2018**, *66*, 2933–2946. [CrossRef]
9. Liu, L.; Yu, W. Massive Connectivity with Massive MIMO—Part II: Achievable Rate Characterization. *IEEE Trans. Signal Process.* **2018**, *66*, 2947–2959. [CrossRef]

10. Wang, Q.; Liu, M.; Liu, N.; Zhong, Z. On Augmenting UL Connections in Massive MIMO System using Composite Channel Estimation. In Proceedings of the IEEE Global Communications Conference (Globecom), Abu Dhabi, UAE, 9–13 December 2018.
11. Zhou, X.; Maham, B.; Hjorungnes, A. Pilot Contamination for Active Eavesdropping. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 903–907. [CrossRef]
12. Kapetanovic, D.; Zheng, G.; Rusek, F. Physical layer security for massive MIMO: An overview on passive eavesdropping and active attacks. *IEEE Commun. Mag.* **2015**, *53*, 21–27. [CrossRef]
13. Lu, Y.; Xiong, K.; Fan, P.; Zhong, Z.; Letaief, K.B. Coordinated Beamforming With Artificial Noise for Secure SWIPT Under Non-Linear EH Model: Centralized and Distributed Designs. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 1544–1563. [CrossRef]
14. Lu, Y.; Xiong, K.; Fan, P.; Zhong, Z.; Letaief, K.B. Robust Transmit Beamforming With Artificial Redundant Signals for Secure SWIPT System Under Non-Linear EH Model. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 2218–2232. [CrossRef]
15. Lu, Y.; Xiong, K.; Fan, P.; Ding, Z.; Zhong, Z.; Letaief, K.B. Global Energy Efficiency in Secure MISO SWIPT Systems With Non-Linear Power-Splitting EH Model. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 216–232. [CrossRef]
16. Kapetanovic, D.; Zheng, G.; Wong, K.K.; Ottersten, B. Detection of pilot contamination attack using random training and massive MIMO. In Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), London, UK, 8–11 September 2013; pp. 13–18.
17. Couillet, R.; Silverstein, J.W.; Bai, Z.; Debbah, M. Eigen-Inference for Energy Estimation of Multiple Sources. *IEEE Trans. Inf. Theory* **2011**, *57*, 2420–2439. [CrossRef]
18. Kapetanovic, D.; Al-Nahari, A.; Stojanovic, A.; Rusek, F. Detection of active eavesdroppers in massive MIMO. In Proceedings of the 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), Washington, DC, USA, 2–5 September 2014; pp. 585–589.
19. Nadler, B. Nonparametric Detection of Signals by Information Theoretic Criteria: Performance Analysis and an Improved Estimator. *IEEE Trans. Signal Process.* **2010**, *58*, 2746–2756. [CrossRef]
20. Kang, J.; In, C.; Kim, H. Detection of Pilot Contamination Attack for Multi-Antenna Based Secrecy Systems. In Proceedings of the IEEE 81st Vehicular Technology Conference (VTC2015-Spring), Glasgow, UK, 11–14 May 2015; pp. 1–5.
21. Yin, H.; Gesbert, D.; Filippou, M.; Liu, Y. A Coordinated Approach to Channel Estimation in Large-Scale Multiple-Antenna Systems. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 264–273. [CrossRef]
22. Flordelis, J.; Gao, X.; Dahman, G.; Rusek, F.; Edfors, O.; Tufvesson, F. Spatial separation of closely-spaced users in measured massive multi-user MIMO channels. In Proceedings of the 2015 IEEE International Conference on Communication (ICC), London, UK, 8–12 June 2015; pp. 1441–1446.
23. Xiong, Q.; Liang, Y.C.; Li, K.H.; Gong, Y. An Energy-Ratio Based Approach for Detecting Pilot Spoofing Attack in Multiple-Antenna Systems. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 932–940. [CrossRef]
24. Tugnait, J.K. Self-Contamination for Detection of Pilot Contamination Attack in Multiple Antenna Systems. *IEEE Wirel. Commun. Lett.* **2015**, *4*, 525–528. [CrossRef]
25. Wang, X.; Liu, M.; Wang, D.; Zhong, C. Pilot Contamination Attack Detection Using Random Symbols for Massive MIMO Systems. In Proceedings of the IEEE 85th Vehicular Technology Conference (VTC2017-Spring), Sydney, Australia, 4–7 June 2017; pp. 1–7.
26. Rissanen, J. Modeling by the shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
27. Rissanen, J. *Information and Complexity in Statistical Modeling*; Publications of the American Statistical Association; Springer: New York, NY, USA, 2007; pp. 1321–1322.
28. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
29. Wax, M.; Kailath, T. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 387–392. [CrossRef]
30. Tulino, A.M.; Verdú, S. *Random Matrix Theory and Wireless Communications*; Now Publishers, Inc.: Hanover, USA, 2004; pp. 1–182.
31. Couillet, R.; Debbah, M. *Random Matrix Methods for Wireless Communications*; Cambridge University Press: Cambridge, UK, 2012.

32. Coon, J.; Sandell, M.; Beach, M.; McGeehan, J. Channel and Noise Variance Estimation and Tracking Algorithms for Unique-word Based Single-carrier Systems. *IEEE Trans. Wirel. Commun.* **2006**, *5*, 1488–1496. [CrossRef]
33. Das, A.; Rao, B.D. SNR and Noise Variance Estimation for MIMO Systems. *IEEE Trans. Signal Process.* **2012**, *60*, 3929–3941. [CrossRef]