# Dempster-Schaffer Theory for Data Fusion in Communication Networks

Prof. D.J.Parish and F. Apperio-Navarro,

Loughborough University (School of Electronic, Electrical and Systems

Engineering)

## Learning Outcomes

1. To understand how Dempster Schafer theory can fuse beliefs in anomaly from multiple sensors..

2. To investigate how DS can be used in anomaly detection in wireless networks

# Dempster-Schafer Theory

## INTRODUCTION

This topic contains the following sections:

- Introduction
- Why use data fusion in network anomaly detection?
- Dempster-Schafer Theory.
- Worked Examples.
- Basic Probability Assignment.
- A Practical Example.

# INTRODUCTION

## Why use Data Fusion?

Statistical anomaly detection processes can never guarantee perfect performance. A specific approach which well in one scenario may fail in a different situation. This suggests that improved overall performance may be achieved if multiple independent detection processes could be used simultaneously. This would produce multiple prediction of anomaly. These need to be combined in an intelligent manner; simply selecting the greatest belief, or calculating an average is unlikely to produce sensible results.

# Bayesian Theorem Mathematic Framework

This mathematical discipline provides the probability of an event $A$ to be true, given that certain evidences $E$ are already known. This is known as conditional probability. The required evidences to calculate the conditional probability are extracted from previous events, occurring under similar experimental conditions to the event $A$. The events are mutually exclusive states of a system. This means that the system can be in only one of these states at a time. The conditional probability provided by the Bayesian Theorem, also known as Posterior probability, is written as in Equation 1:

$$P(A|E) = \frac{P(E|A)\,P(A)}{P(E)} = \frac{P(E|A)\,P(A)}{[P(A)\,P(E|A) + P(\bar{A})\,P(E|\bar{A})]} \qquad (6.1)$$

According to this definition, Bayesian theory is unable to assign probability in the considered event in the absence of any other knowledge. Only after evidence *E* is obtained, can the posterior probability be computed. From the previous equation, three terms can be described. The term $P(A)$ reflects the probability that a particular event is true in the absence of evidence. This is generally known as the a priori probability. The a priori probability,$P(A)$, would be updated along with the posterior repetition of the consider event $A$.

## DEMPSTER-SCHAFER THEORY

Dempster-Shafer theory of evidence method is a discipline of mathematics that combines evidence of information from multiple and heterogeneous events in order to calculate the probability of occurrence of another event.

The D-S theory starts by assuming a Universe of Discourse $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$, also called a Frame of Discernment, which is a finite set of all possible mutually exclusive propositions and hypotheses about some problem domain.

With regards to anomaly based network attack detection, the frame of discernment is comprised of *A* = "*Attack*" and *N* = "*Normal*". Assuming Θ has two outcomes {*A, N*}, the total number of subsets of Θ, defined by the number of hypotheses that it composes, is $2^{\Theta}$ = {*A, N*, {*A*|*N*}, Ø}

Each proposition (subset) from Θ is assigned a probability or a confidence interval within [0, 1], by an observer from the mass probability function $m$, also known as the basic probability assignment:

$$m: 2^{\Theta} \to [0,1] \qquad if \begin{cases} m(\emptyset) = 0 \\ m(A) \geq 0, \forall A \subseteq \Theta \\ \sum_{A \subseteq \Theta} m(A) = 1 \end{cases}$$

The function $m(A)$ is defined as *A*'s basic probability number. It describes the measure of belief that is committed exactly to hypothesis *A*.

In order to define the confidence interval that is given to a certain event, two functions must first be defined. These are the Belief function ($Bel$) and the Plausibility function ($Pl$). The former is a belief measure of a hypothesis *A*, and it sums the mass value of all the non-empty subsets of *A*.

$$Bel(A) = \sum_{B \subseteq A} m(B) \qquad \forall A \subseteq \Theta$$

The doubt function ($Dou$) is given by

$$Dou(A) = Bel(\neg A) = 1 - \sum_{B \cap A = \emptyset} m(B)$$

which accounts for all evidence that rule out the given proposition represented by *A*.

Similarly, the $Pl$ function takes into account all the evidence that does not rule out the given proposition. In other words, it expresses how much we should believe in *A* if all currently unknown facts were to support *A*.

$$Pl(A) = 1 - Dou(A)$$

Thus, the true belief in hypothesis *A* will be along the interval $[Bel(A), Pl(A)]$. However, in practice, the values of the interval could be identical and therefore the interval becomes a unique value.

The idea behind the D-S rule of combination is to fuse the belief from two different observers into one given hypothesis.

EVENT PROBABILITIES ASSIGNED BY $m_1$ AND $m_2$

| $m_2 \backslash m_1$ | $\{A\}$: 0.32 | $\{N\}$: 0.25 | $\{A, N\}$: 0.43 |
|---|---|---|---|
| $\{A\}$: 0.35 | 0.11 | 0.09 | 0.15 |
| $\{N\}$: 0.1 | 0.03 | 0.025 | 0.04 |
| $\{A, N\}$: 0.55 | 0.18 | 0.14 | 0.24 |

Let $m_1$ and $m_2$ be the basic probability assignments from observer 1 and 2 respectively. The cells in the above table represent the multiplication of the $m_1$ belief with the $m_2$ belief, horizontal and vertical axis, respectively.

Their orthogonal $sum, m = m_1 \oplus m_2$, is defined as

$$m(A) = \frac{\sum_{X\cap Y=A} m_1(X) * m_2(Y)}{1 - \sum_{X\cap Y=\emptyset} m_1(X) * m_2(Y)} \quad \forall A \neq \emptyset \qquad (1)$$

If the denominator of eq. (1) is equal to zero, $K = 0$, then $m_1 \oplus m_2$ does not exist and $m_1$ and $m_2$ are said to be totally or flatly contradictory.

To easily understand how to apply the D-S algorithm, a real example from our measurements is presented. The basic probabilities for an event being "*Attack*", "*Normal*", and "*Uncertain*", can be tabulated as seen in Table I.

Firstly *K* is calculated from eq. (1): $K = 1 - (0.03 + 0.09) = 0.88$. As described in eq. (1), for any event E the combined belief is given by:

$$m(E) = \frac{1}{K} * \sum_{X\cap Y=E} m_1(X) * m_2(Y)$$

Therefore,

$$m(A) = 0.88 * (0.11 + 0.15 + 0.18) = 0.39$$
$$m(N) = 0.88 * (0.025 + 0.04 + 0.14) = 0.18$$
$$m(A|N) = 0.88 * (0.24) = 0.21$$

According to the results, the hypothesis more likely to be true is *A*, with higher belief than the other hypotheses.

### An Example for Three Sensors

To easily understand how to apply the D-S theory, an example using three sensors is presented here.

Let us consider one system with three sensors, Sensor 1, Sensor 2 and Sensor 3. These sensors monitor and gather frames from a WLAN. Using the combined evidences of information provided by the three sensors, the system needs to classify the gathered frames either as malicious or non-malicious.

In such scenario, the frame of discernment is comprised of two possible outcomes, $A = Attack$ and $N = Normal$. Hence, the total number of hypotheses considered for this example would be $2^{\Theta} = \{A, N, \{A|N\}, \emptyset\}$. Each sensor provides an independent belief in each possible hypothesis. The beliefs assigned by the three sensors are combined to calculate a final decision; i.e. whether the gathered frames are malicious or not. The basic probabilities for one of the frames is tabulated in the table below.

EVENT PROBABILITIES ASSIGNED BY $m_1$ (HORIZONTAL X) AND $m_2$ (VERTICAL Y).

| $m_2$ \ $m_1$ | $m_1(N) = 0.25$ | $m_1(A) = 0.32$ | $m_1(A\|N) = 0.43$ | $m_1(\emptyset) = 0$ |
|---|---|---|---|---|
| $m_2(N) = 0.1$ | 0.025 | 0.032 | 0.043 | 0 |
| $m_2(A) = 0.35$ | 0.0875 | 0.112 | 0.1505 | 0 |
| $m_2(A\|N) = 0.55$ | 0.1375 | 0.176 | 0.2365 | 0 |
| $m_2(\emptyset) = 0$ | 0 | 0 | 0 | 0 |

The horizontal axis of the Table 4.1. represents the beliefs of the Sensor 1, for each hypothesis. Similarly, the vertical axis represents the beliefs of the Sensor 2, for all the hypotheses. The cells in the table represent the multiplication of the beliefs of both sensors.

Dempster's rule of combination is used to combine the beliefs and generate a final decision. The results for the first iteration of this example are:

$$m_{12}(N) = \frac{(0.025 + 0.1375 + 0.043)}{1 - (0.0875 + 0.032)} = 0.233$$

$$m_{12}(A) = \frac{(0.112 + 0.1505 + 0.176)}{1 - (0.0875 + 0.032)} = 0.498$$

$$m_{12}(A|N) = \frac{(0.2365)}{1 - (0.0875 + 0.032)} = 0.269$$

Then, the output results of this initial combination process are used as input evidences in the next iteration, along with the evidences of information from the Sensor 3. The horizontal axis of the table below. represents the beliefs of the Sensor 3, for each hypothesis. Similarly, the vertical axis represents the combined beliefs of the Sensors 1 and 2, for all the hypotheses.

EVENT PROBABILITIES ASSIGNED BY $m_3$ (HORIZONTAL X) AND $m_{12}$ (VERTICAL Y).

| $m_{12}$ \ $m_3$ | $m_3(N) = 0.27$ | $m_3(A) = 0.41$ | $m_3(A\|N) = 0.32$ | $m_3(\emptyset) = 0$ |
|---|---|---|---|---|
| $m_{12}(N) = 0.233$ | 0.063 | 0.0955 | 0.0745 | 0 |
| $m_{12}(A) = 0.498$ | 0.134 | 0.2045 | 0.1595 | 0 |
| $m_{12}(A\|N) = 0.269$ | 0.073 | 0.11 | 0.086 | 0 |
| $m_{12}(\emptyset) = 0$ | 0 | 0 | 0 | 0 |

The results for this iteration are:

$$m(N) = \frac{(0.063 + 0.073 + 0.0745)}{1 - (0.134 + 0.0955)} = 0.273$$

$$m(A) = \frac{(0.2045 + 0.11 + 0.1595)}{1 - (0.134 + 0.0955)} = 0.615$$

$$m(A|N) = \frac{(0.086)}{1 - (0.134 + 0.0955)} = 0.112$$

According to the combined results of the evidences of information from the three sensors, the belief in the hypothesis $A$ is higher than the other two hypotheses. Therefore, the hypothesis more likely to be true is $A$, with 61.5% of belief in $Attack$.

**Further Considerations**

D-S is suitable for detecting previously unseen attacks because it does not require a priori knowledge. Also, it provides the ability of managing and assigning probability to ignorance, which allows it to tackle a large range of problems.

In contrast, Bayesian inference requires a priori knowledge and does not allow allocation of probability to ignorance but only to an event being normal or abnormal.

Nevertheless, there are two main drawbacks associated with the D-S algorithm. First, the high computation complexity and second the conflicting beliefs management. The computational complexity increases exponentially with the number of possible event outcomes (Θ). If there are $n$ elements in Θ, there will be up to $2^n$ - 1 focal elements for the mass functions, ignoring Ø. The combination of two mass functions needs the computation of up to $2^n$ intersections.

The frame of discernment in the proposed methodology includes two elements ($n = 2$), normal and abnormal, and therefore there will be three focal elements of belief functions, $2^2$ = {*Attack*, *Normal*, {*Attack* | *Normal*}, Ø}. By using only three elements in the focal elements, the fusion method requires low computational complexity.

The conflicting belief phenomenon is nicely illustrated with a simple example. Given three events, {*A*, *B*, *C*} and two sensors, Sensor 1 might assign $m(A) = 0.9$, $m(B) = 0.1$ and $m(C) = 0$ as beliefs in *A*, *B* and *C* respectively. Similarly, Sensor 2 might assign $m(A) = 0$, $m(B) = 0.1$ and $m(C) = 0.9$ as beliefs in *A*, *B* and *C*. Applying the D-S algorithm on these values, the rule of combination will result with a higher belief in event *B*, which is clearly wrong. In the proposed detection algorithm of this work, each event is assigned a non-zero mass function and therefore the belief conflict phenomenon is not an issue.

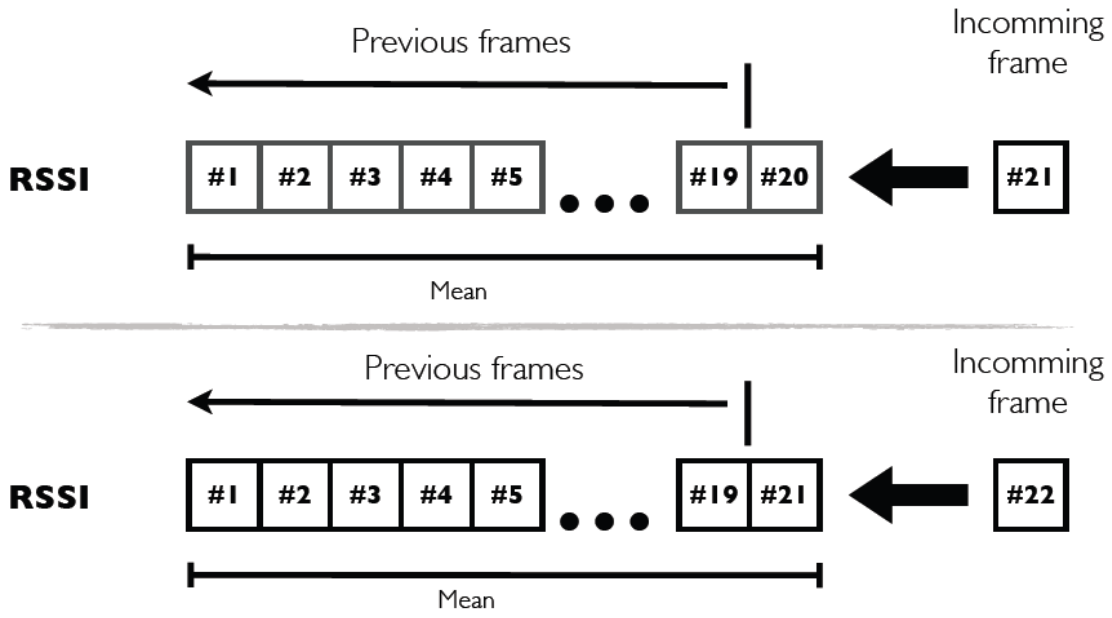
Another issue relates to the independence of the beliefs used. This is usually impossible to guarantee in a communication network application. Research however indicates that the technique can still be applied where a degree of correlation is present. In some cases, this can be accommodated by adjusting the belief values.

## Basic Probability Assignment

In the IDS literature there exist multiple ways of assigning probabilities to each of the hypotheses in D-S theory, ranging from data mining techniques to empirical approaches. For instance, expert opinion may be utilised to manually assign the belief probabilities to each of the hypotheses. This BPA process is completely subjective and might not be adequate for automatic and self-adaptive IDSs. An alternative approach is a methodology that seeks changes in the signal-to-noise ratio. The value of this single metric is measured from distinct nodes running two different local algorithms, single threshold and cumulative sum. Based on the measured information, their system generates the BPAs through the use of a linear function. One of the drawbacks of this methodology is that both local algorithms require the utilisation of diverse tuning parameters. In yet another approach, multiple manually defined thresholds are used, empirically defined after analysing non-malicious data. The authors do not describe the way the thresholds are defined. One last example uses data mining techniques to proceed with the BPA tasks.

### A lightweight BPA assignment process

This is based on a simple sliding window.




The sliding window scheme works as follows. The first time the IDS is run, the $n$ slots within the sliding window will be initially filled with frames metrics before being able to detecting intrusions. Once the $n$ slots within the sliding window have been filled, each of the $n$ frames metrics is analysed and the reference of normality is generated. After all the frames within the first sliding window have been analysed and the detection implemented, the system slides the window one single slot. The metric from the next incoming frame is stored in the slot that becomes empty after sliding the window. Then, a new reference of normality is calculated using the previous $(n-1)$ frames along with the last stored frame. After the new reference of normality has been calculated, only the last stored frame is analysed, since the previous $(n-1)$ frames have already been analysed. Next, the system slides the window one single slot again and a new frame is included. The described process is constantly repeated. This configuration allows detecting attacks as they occur.

The methodology that has been proposed to assign beliefs in $Normal$ is based on the degree of dispersion of the values in the dataset. We consider the total number of instances in the dataset ($n$), the first quartile ($Q_1$) that defines the boundary for the lower 25% of the data, the second quartile, or median ($Me$), that defines the boundary for the 50% of the data, and the third quartile ($Q_3$) that defines the boundary for the lower 75% of the data. To calculate these three parameters, the $n$ instances in the dataset are sorted from the lowest to highest value. The $Me$ is the data instance that, after being sorted, divides the dataset in half, leaving the lowest 50% of the dataset at one side and the highest 50% at the other side. The $Q_1$ is the data instance that, after being sorted, leaves the lowest 25% of the dataset at one side and the highest 75% at the other side. The $Q_3$ is the data instance that, after being sorted, leaves the lowest 75% of the dataset at one side and the highest 25% at the other side. Also, the interquartile range ($IQR$), the

difference between $Q_3$ and $Q_1$ represented in Equation (6.6), as well as the $Min$ and $Max$ values are calculated. These two last parameters are calculated using the following Equations.

$$Min = Q_1 - 1.5 \times IQR \quad (6.6)$$

$$Max = Q_3 + 1.5 \times IQR \quad (6.7)$$

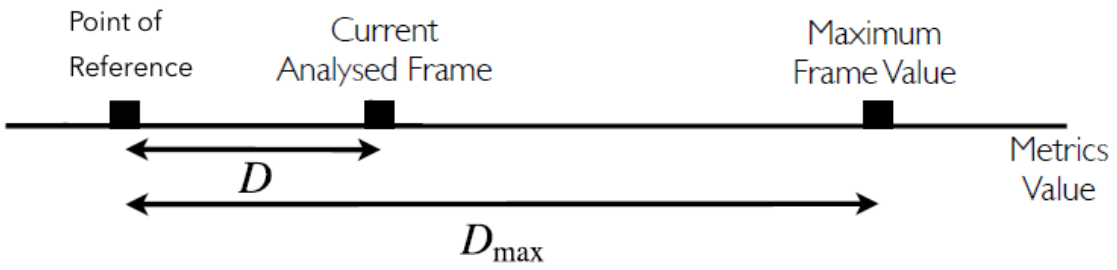
$$IQR = Q_3 - Q_1 \quad (6.8)$$

The methodology that has been proposed to assign beliefs in $Attack$ is based on the distance from the currently analysed instance to a point of reference. It is necessary to start by defining a certain number of parameters. Again, $n$ represents the total number of instances in the dataset. It is required to identify a point of reference, as well as the data instance with the highest value ($Hi$) and the instance with the lowest value ($Lo$). After sorting the $n$ instances in the dataset, it is straightforward to select $Hi$ and $Lo$. The parameter $M$ is calculated using the Equation below.

$$M = \frac{1}{n}\sum_{i=1}^{n} a_i \quad (6.9)$$

where $n$ is the total number of instances.

Once the point of reference is selected, the Euclidean distances from the point of reference to the lowest value ($Lo$) and the highest value ($Hi$) are calculated. The value with the largest Euclidean distance ($D_{max}$) from the point of reference represents the maximum possible belief in the hypothesis $Attack$. Next, the Euclidean distance from the point of reference to the currently analysed data instance ($D$) is also calculated. Finally, the belief in $Attack$ is assigned using a simple linear function, making use of the different parameters calculated.




. The $Uncertainty$ has been considered as an adjustment parameter to satisfy the required conditions of D-S theory.

The outcome of the two previous methods could provide four different and mutually exclusive situations:

- Low belief in $Attack$ and high belief in $Normal$.
- High belief in $Attack$ and low belief in $Normal$.
- High belief in $Attack$ and high belief in $Normal$.
- Low belief in $Attack$ and low belief in $Normal$.

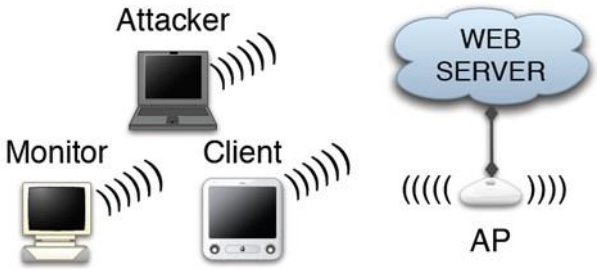
The methodology to assign the belief in $Uncertainty$ normalises the smaller of the other two beliefs ($Belief_{Min}$) to the largest ($Belief_{Max}$). In line with the previous two methodologies, the maximum BPA value has been limited to 50%. The belief in $Uncertainty$ is calculated using the Equation below.

$$Belief_{Unc.} = \frac{0.5 * Belief_{Min}}{Belief_{Max}}$$

# Using the Theory in a real Detection System

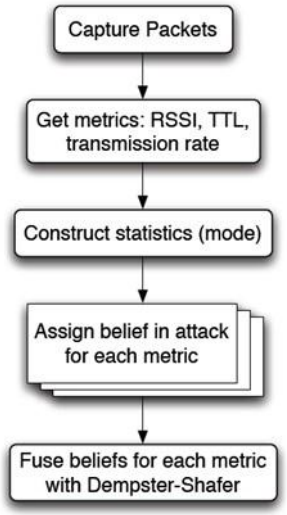
Dempster-Schaffer Theory can be used to enhance the success of an anomaly based detection approach by fusing multiple beliefs about an event. These multiple beliefs could come from multiple sensors (perhaps seeing traffic at different locations in a network), or from different metrics related to traffic seen at one point in the network. (Note that D-S Theory assumes that the beliefs are not correlated so this can restrict their choice). An important benefit of using D-S for this task is the ability of the Theory to assign a weighting to uncertainty.

The system to be described collects packets (frames) from an IEEE 802.11 (WiFi) network and looks for changes in the basic statistical characteristics of a section of packet/frame related metrics using a simple window. A prediction of Attack/No Attack is then made for each metric and these are fused using a D-S system.




**TestBed Example for a Man-in-the-Middle Attack.**

## Software Procedure




## Results for Different Data.

The results below show the detection performance (False Negatives and False Positives) for different metrics and combinations. Better results are generally seen for increasing numbers of metrics (upto 3

in this example).

SINGLE METRIC RESULTS UTILISING TTL

| Web Site | Type | OSR (%) | False Neg. (%) | False Pos. (%) |
|---|---|---|---|---|
| China | Normal | 100 | 0 | 4.06 |
| | Attack | 100 | 0 | 2.74 |
| Spain | Normal | 100 | 0 | 5.24 |
| | Attack | 100 | 0 | 5.22 |
| UK | Normal | 100 | 0 | 14.58 |
| | Attack | 97.50 | 2.50 | 20.73 |
| US 01 | Normal | 100 | 0 | 9.60 |
| | Attack | 97.37 | 2.63 | 6.67 |
| US 02 | Normal | 100 | 0 | 22.26 |
| | Attack | 87.32 | 12.68 | 12.42 |

DUAL METRIC RESULTS UTILISING INJ. RATE AND RSSI

| Web Site | Type | OSR (%) | False Neg. (%) | False Pos. (%) |
|---|---|---|---|---|
| China | Normal | 100 | 0 | 0 |
| | Attack | 100 | 0 | 0.24 |
| Spain | Normal | 100 | 0 | 0.35 |
| | Attack | 73.33 | 26.67 | 3.73 |
| UK | Normal | 100 | 0 | 0.49 |
| | Attack | 73.17 | 26.83 | 2.59 |
| US 01 | Normal | 100 | 0 | 0 |
| | Attack | 100 | 0 | 0.75 |
| US 02 | Normal | 100 | 0 | 0.29 |
| | Attack | 91.84 | 8.16 | 1.22 |

DUAL METRIC RESULTS UTILISING INJ. RATE AND TTL

| Web Site | Type | OSR (%) | False Neg. (%) | False Pos. (%) |
|---|---|---|---|---|
| China | Normal | 100 | 0 | 0.04 |
| | Attack | 100 | 0 | 0.43 |
| Spain | Normal | 100 | 0 | 0 |
| | Attack | 93.33 | 6.67 | 0.77 |
| UK | Normal | 100 | 0 | 1.17 |
| | Attack | 95.62 | 4.38 | 1.78 |
| US 01 | Normal | 100 | 0 | 0.27 |
| | Attack | 82.35 | 17.65 | 3.85 |
| US 02 | Normal | 100 | 0 | 2.33 |
| | Attack | 78.74 | 21.26 | 5.76 |

CROSS LAYER RESULTS UTILISING RSSI, INJ. RATE AND TTL

| Web Site | Type | OSR (%) | False Neg. (%) | False Pos. (%) |
|---|---|---|---|---|
| China | Normal | 100 | 0 | 0 |
| | Attack | 100 | 0 | 0 |
| Spain | Normal | 100 | 0 | 0 |
| | Attack | 100 | 0 | 0 |
| UK | Normal | 100 | 0 | 0 |
| | Attack | 90.45 | 9.55 | 4.70 |
| US 01 | Normal | 100 | 0 | 0 |
| | Attack | 85.71 | 14.29 | 3.71 |
| US 02 | Normal | 100 | 0 | 0 |
| | Attack | 100 | 0 | 0.08 |