Table 5: A portion of refined correlations acquired from Testbed 1.

ID

Correlation

ID

Correlation

ID

Correlation

ID

Correlation

C1

C2

remotion(MS1)

C3

presence(PS1)

contact(Cl)

C4

/ppresence(PS1)

C5

C6

C7

presence(PS2)

C8

C9

C10

C11

C12

C13

(Eswitch(L4)

C14

(Eswitch(L3)

C15

switch(L3)

C16

(Fswitch(P2)

C17

/qacceleration(C3)

C18

C19

switch(L4)

C20

/c-ontact(C1)

Eclosed

C21

(contact(C3)

C22

mgtion(MS3)

C23

C24

".illuminance(MS1)

C25

/presence(PS1)

C26

motion(MS1)

C27

C28

C29

C30

C31

C32

C33

C34

C35

C36

C37

1)cep

C38

Table 6: Impact of Different Training-Phase Duration

one-tail test (Section 5.3.3), which has two impacts. On the

Training phase

Precision

Recall

# of false alarms

# of correlations

(days)

one hand, even a very small number of abnormal behaviors

3

63.63%

78.69%

212

183

in the small datasets will cause some true correlations to be

6

75.35%

85.78%

147

141

9

94,57%

94,12%

15

135

rejected. On the other hand, due to the small amount of data,

12

97,25%

94.12%

8

132

many false correlations are not rejected yet. (3) Starting from

15

97.83%

94.12%

4

130

the dataset of 15 days, the performance (including the num-

18

97.83%

94,12%

4

130

21

97.83%

94.12%

4

130

ber of false alarms) does not change anymore, which means

that amount of data is sufficient for the testbed. (4) Those true

correlations which have been rejected in the small datasets

quent event set and 214 of them have that in their antecedent

are recovered in the larger datasets. This shows the robust-

event set. There are 80 rules involving lights L4 and L5, 32

ness of the design of HAWatcher. Even if very few anomalies

with illuminance sensors in MS3 and MS4, and 14 with the

arise during the training phase, true correlations can survive

CO2 sensor in A. Other attributes are not seen in any rules,

given sufficient training data. (5) We examine the different

as events involving them are overshadowed by those involv-

sets of correlations mined based on different duration and

ing the four aforementioned attributes. In contrast, with our

find that some false correlations remain there until

mining method, each attribute is involved in at least four (4) data is available. For

correlations and has an average of 10.5 correlations.

remains until behaviors that fail the correlation appear

For the OCSVM-based detector, it takes a snapshot of all

Days 11 and 12.

devices' states as a frame each time a new event arises and

concatenates four consecutive frames as one input data vec-

6.3

Anomaly Generation

tor [48]. We use the open source OCSVM implemetation in

sklearn [63] and the default kernel (Radial Basis Function).

To evaluate HAWatcher, we simulate 24 cases of anomalies

Impact of Training-Phase Duration We study the impact

on Testbed 1 listed in Table 7 (totally 62 cases on the four

of the duration of the training phase on the performance of

testbeds). We follow two criteria to select anomaly cases:

HAWatcher. As Testbed 1 is the most complex one among the

(1) the attacks are discussed in the literature about IoT at-

four testbeds, we select it in this experiment. As illustrated

tacks; and (2) the malfunctions are frequently discussed in

in Table 6, we start from using the first three (3) days of data

the SmartThings community. To simulate an anomaly case,

as a training dataset, and then use the first six (6) days by

we either modify the testing event logs (collected in the

increasing three days of data, and SO on until we use all the

fourth week) or interfere with the home automation, and the

21 days of data. With each of the seven (7) training datasets,

resulting logs are used for anomaly detection. For each case,

we train a system and evaluate its performance using the

multiple instances (see the "#inst." column) are injected.

fourth week of testing data.

If an attack has the same impact on the event logs as a

Based on the study and the results shown in Table 6, we

malfunction, we group and simulate them as one case. Taking

have the following observations. (1) Nine (9) days of training

Case 1 as an example, we randomly inject a total of 50 motion

data is enough for HAWatcher to achieve the highest detec-

events of MS1 into the testing event logs to simulate the effect

tion recall, but its number of false alarms has not reached the

of both Faulty Events (due to sensor malfunctions) and Fake

lowest, which means some false correlations are obtained. (2)

Events (due to attacks).

For the first two training datasets, although they lead to more

Faulty/Fake Events. We simulate them by inserting events

correlations than the subsequent ones, the overall quality

of devices, such as motion sensors [17], presence sensor [14],

of correlations is not high. The reason is that we use the

and contact sensors [3], as they are reportedly unreliable.