US 20230007042A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2023/0007042 A1

Haworth et al. (43) Pub. Date: Jan. 5, 2023

(54) **A METHOD AND SYSTEM FOR DETERMINING AND ACTING ON AN EMAIL CYBER THREAT CAMPAIGN**

(71) Applicant: **Darktrace Holdings Limited**, Cambridge (GB)

(72) Inventors: **Stephen Haworth**, Cambridge (GB); **Stephen Pickman**, Huntingdon (GB); **Antony Steven Lawson**, Tyne & Wear (GB); **Paul Lancaster**, Cambridge (GB)

(21) Appl. No.: **17/859,847**

(22) Filed: **Jul. 7, 2022**

Related U.S. Application Data

(63) Continuation-in-part of application No. 17/187,381, filed on Feb. 26, 2021, which is a continuation-in-part of application No. 16/732,644, filed on Jan. 2, 2020, now Pat. No. 11,477,222, which is a continuation-in-part of application No. 16/278,932, filed on Feb. 19, 2019.

(60) Provisional application No. 63/219,026, filed on Jul. 7, 2021, provisional application No. 63/317,157, filed on Mar. 7, 2022, provisional application No. 62/983,307, filed on Feb. 28, 2020, provisional application No. 63/026,446, filed on May 18, 2020, provisional application No. 62/796,507, filed on Jan. 24, 2019, provisional application No. 62/632,623, filed on Feb. 20, 2018.

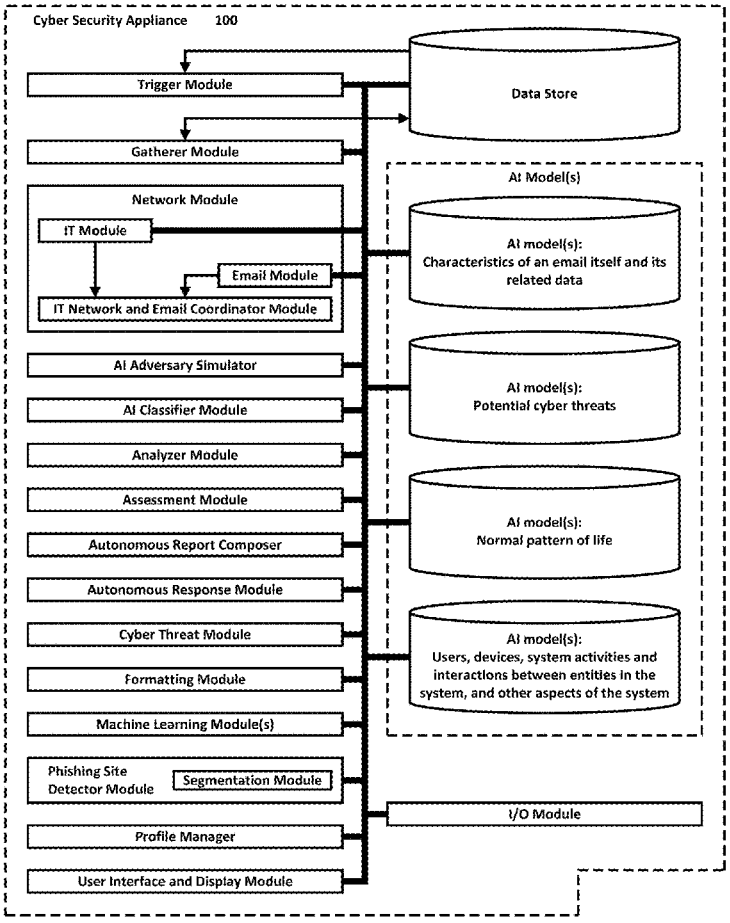
Publication Classification

(51) **Int. Cl.**
H04L 9/40 (2006.01)

(52) **U.S. Cl.**
CPC ***H04L 63/1466*** (2013.01); ***H04L 63/1425*** (2013.01); ***H04L 63/205*** (2013.01)

(57) **ABSTRACT**

A cyber security appliance (CSA) configurable to protect a computer system from email cyber threat campaigns is disclosed. The CSA may comprise: an email module configured to process all incoming emails and log data and metadata; a cyber threat module coupled configured to assess a severity level of a cyber threat using one or more Artificial Intelligence (AI) models; an AI classifier configured to determine the likelihood of an email cyber threat campaign; an autonomous response module configured to act against emails determined to be threats; and a user interface module configured to generate a report, present data on a display, and show a graphical display of the system indicating the details of a cyber threat campaign.



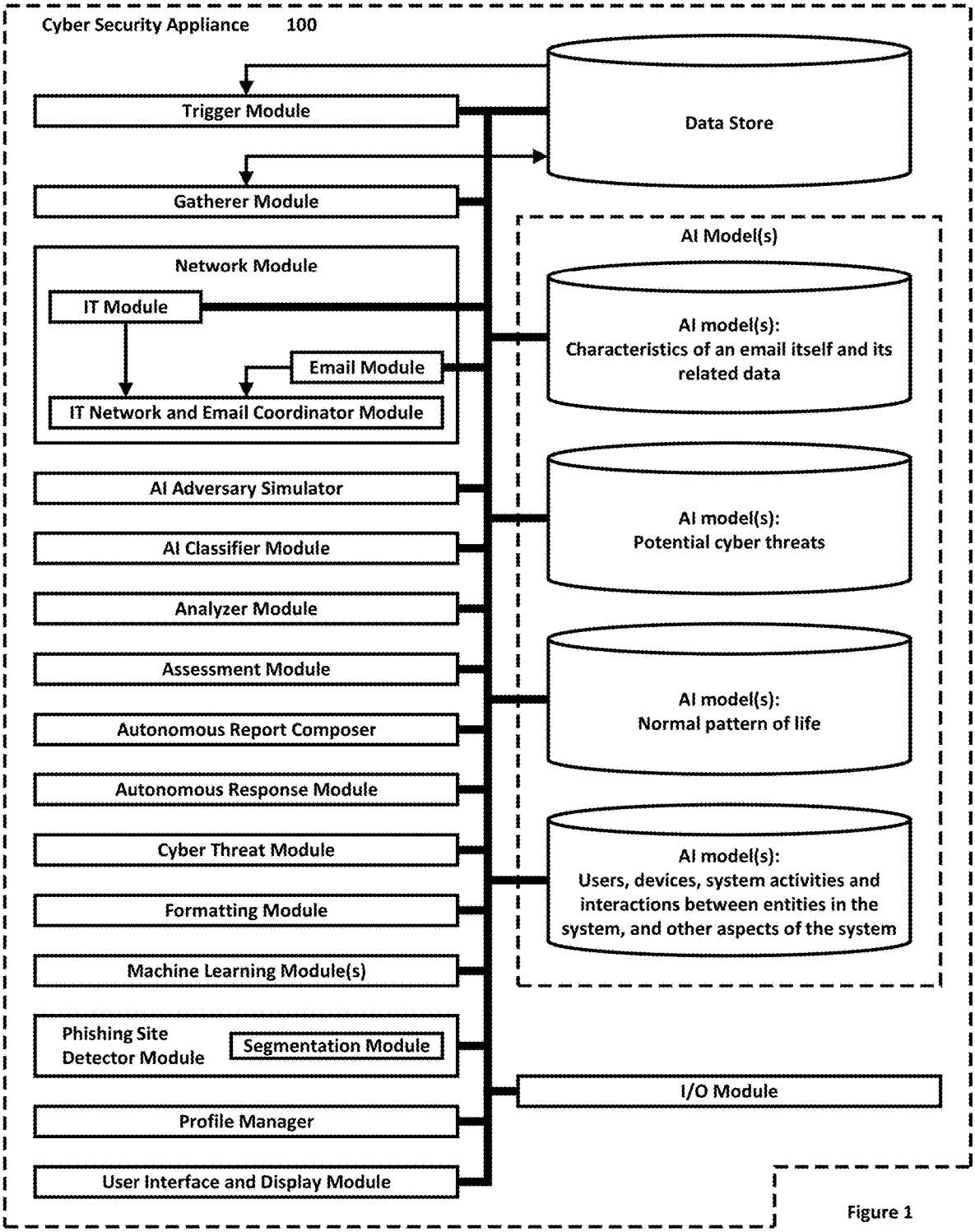

Cyber Security Appliance 100
Trigger Module
Gatherer Module
Network Module
IT Module
Email Module
IT Network and Email Coordinator Module
AI Adversary Simulator
AI Classifier Module
Analyzer Module
Assessment Module
Autonomous Report Composer
Autonomous Response Module
Cyber Threat Module
Formatting Module
Machine Learning Module(s)
Phishing Site Detector Module
Segmentation Module
Profile Manager
User Interface and Display Module
Data Store
AI Model(s)
AI model(s): Characteristics of an email itself and its related data
AI model(s): Potential cyber threats
AI model(s): Normal pattern of life
AI model(s): Users, devices, system activities and interactions between entities in the system, and other aspects of the system
I/O Module

Figure 1

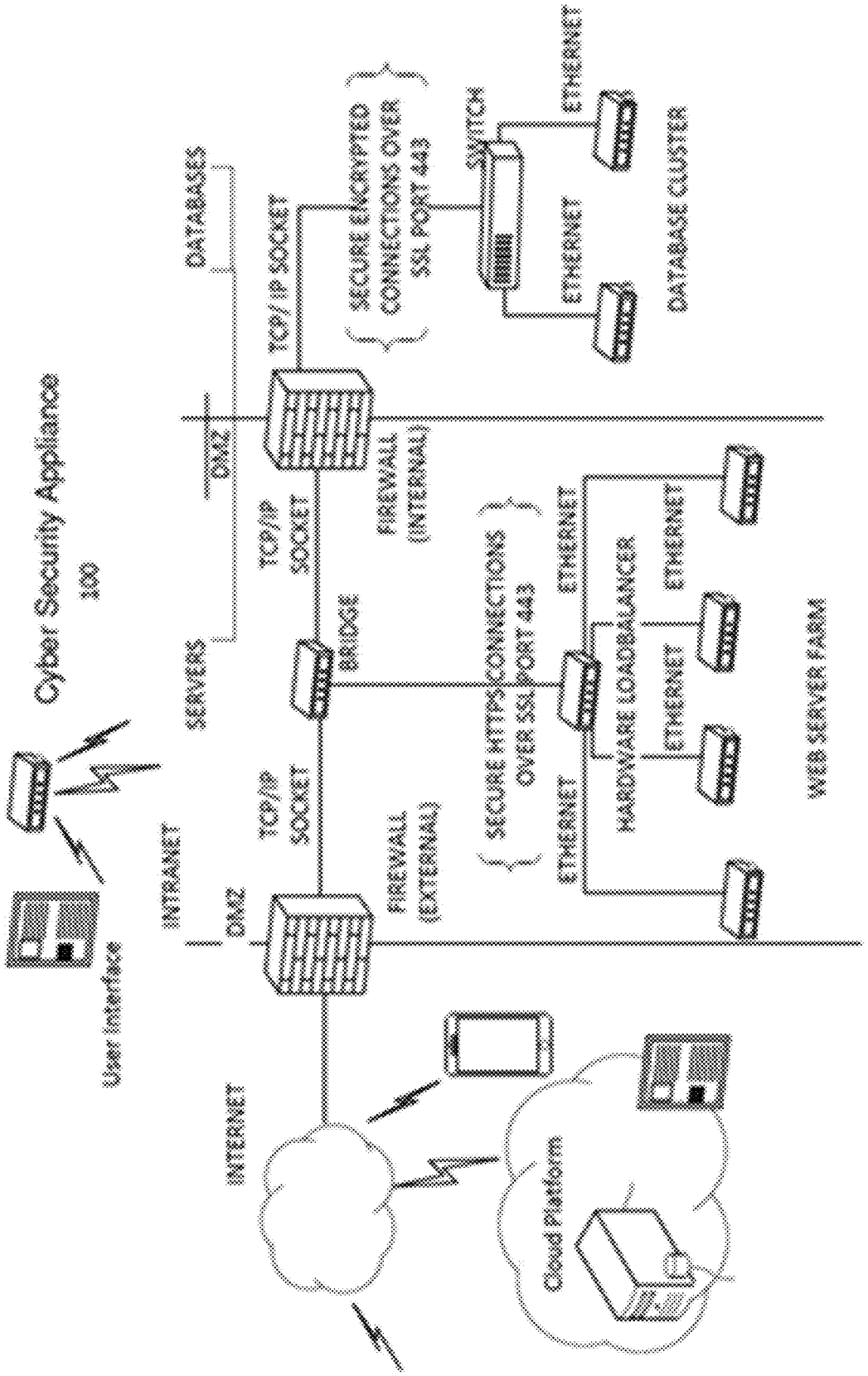

Cyber Security Appliance
100
User Interface
INTRANET
DMZ
TCP/IP SOCKET
FIREWALL (EXTERNAL)
SECURE HTTPS CONNECTIONS OVER SSL PORT 443
ETHERNET
HARDWARE LOADBALANCER
ETHERNET
ETHERNET
WEB SERVER FARM
SERVERS
DMZ
TCP/IP SOCKET
BRIDGE
FIREWALL (INTERNAL)
DATABASES
TCP/IP SOCKET
SECURE ENCRYPTED CONNECTIONS OVER SSL PORT 443
SWITCH
ETHERNET
ETHERNET
DATABASE CLUSTER
INTERNET
Cloud Platform

FIGURE 2

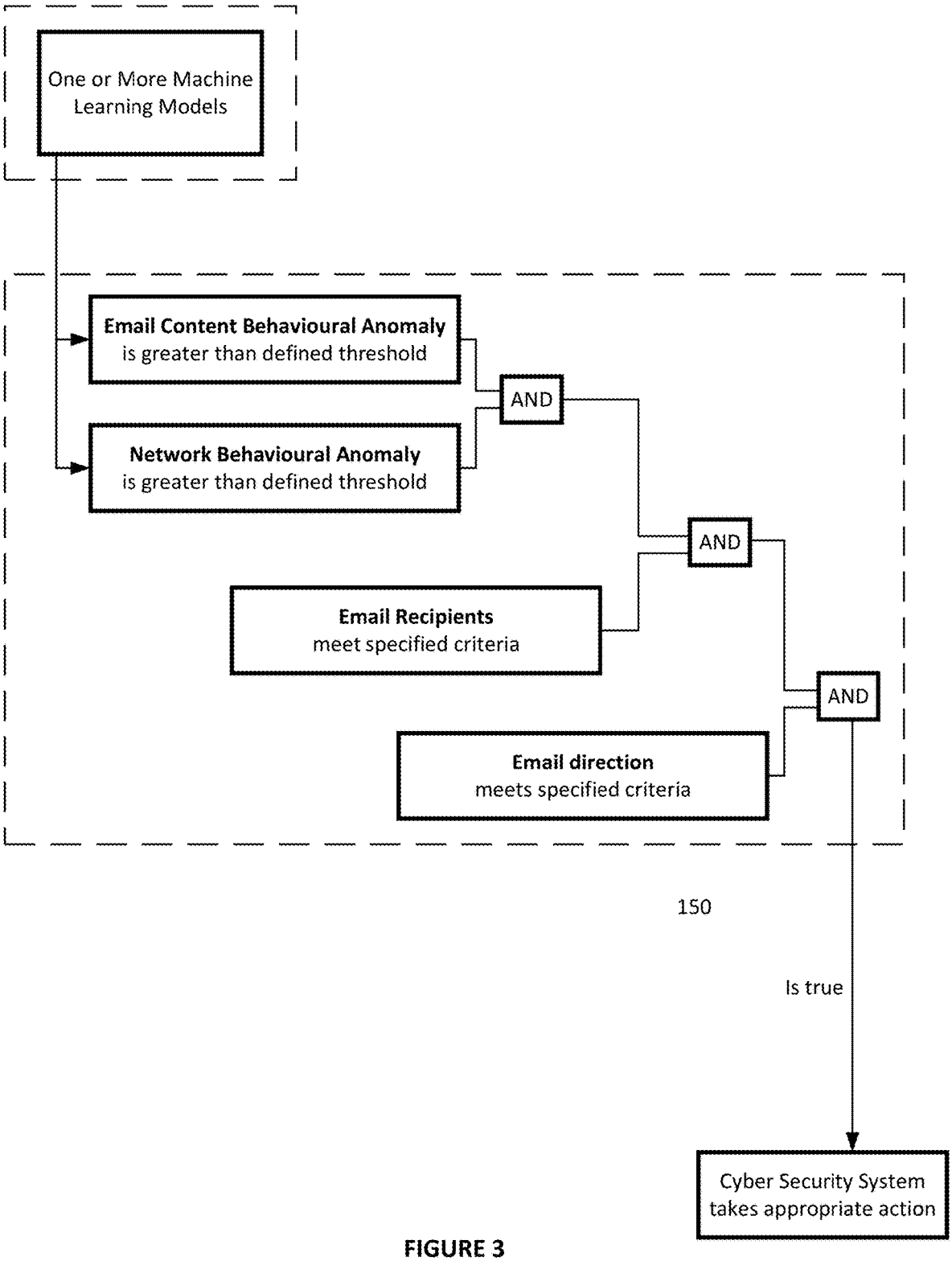

One or More Machine Learning Models
Email Content Behavioural Anomaly is greater than defined threshold
Network Behavioural Anomaly is greater than defined threshold
AND
AND
AND
Email Recipients meet specified criteria
Email direction meets specified criteria
150
Is true
Cyber Security System takes appropriate action

FIGURE 3

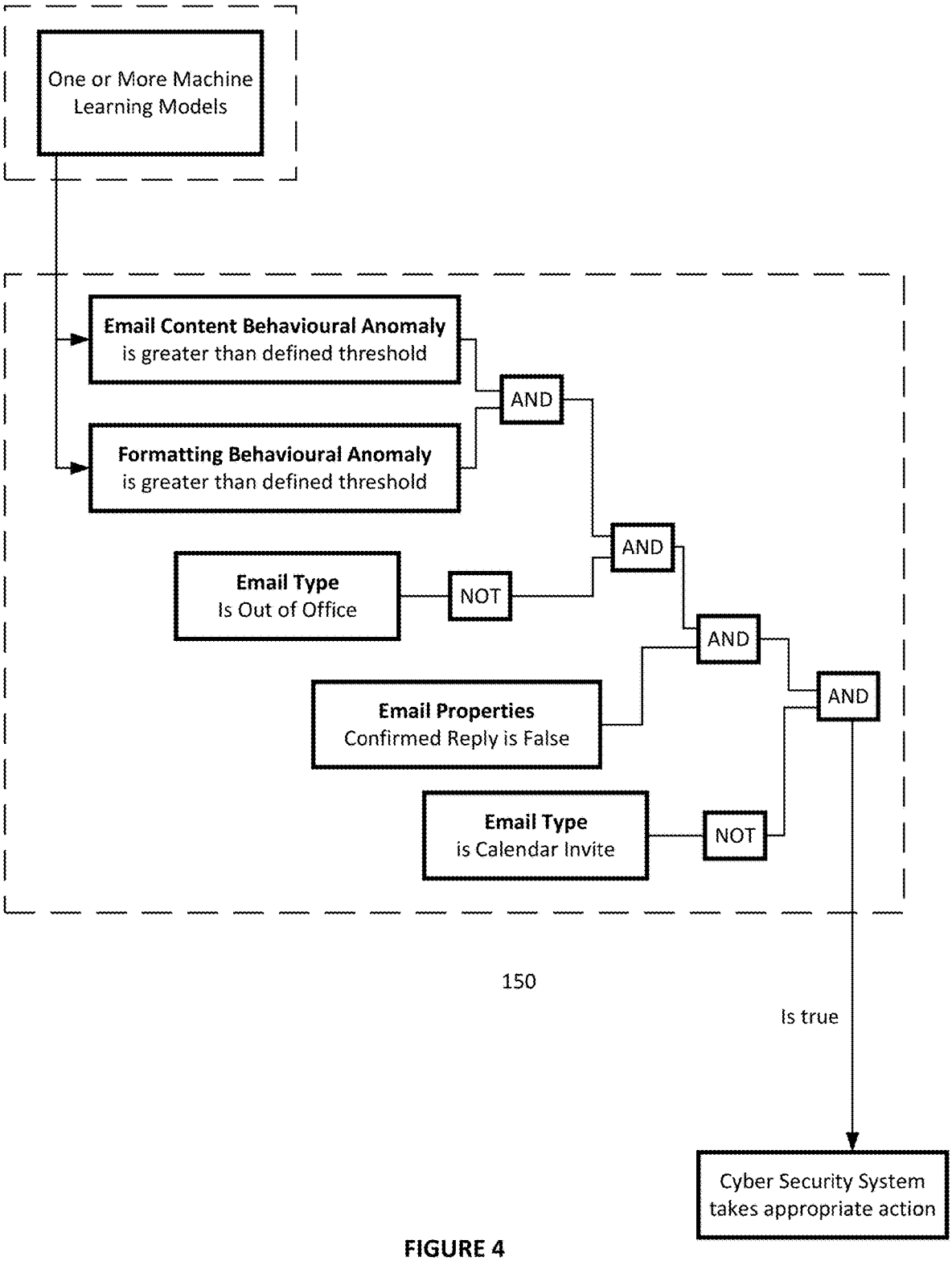

One or More Machine Learning Models
Email Content Behavioural Anomaly is greater than defined threshold
Formatting Behavioural Anomaly is greater than defined threshold
AND
Email Type Is Out of Office
NOT
AND
AND
Email Properties Confirmed Reply is False
Email Type is Calendar Invite
NOT
AND
150
Is true
Cyber Security System takes appropriate action

FIGURE 4

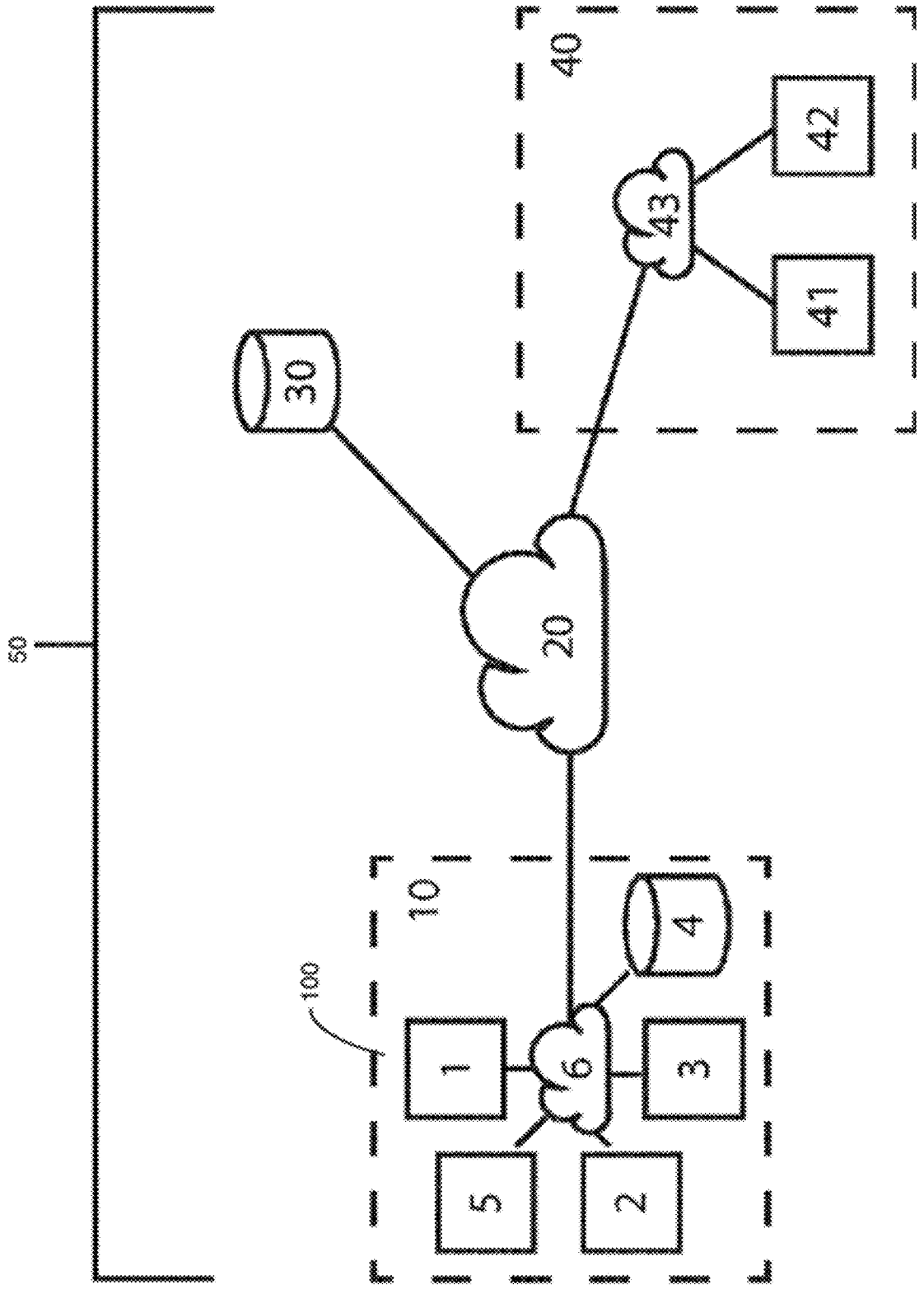


FIGURE 5

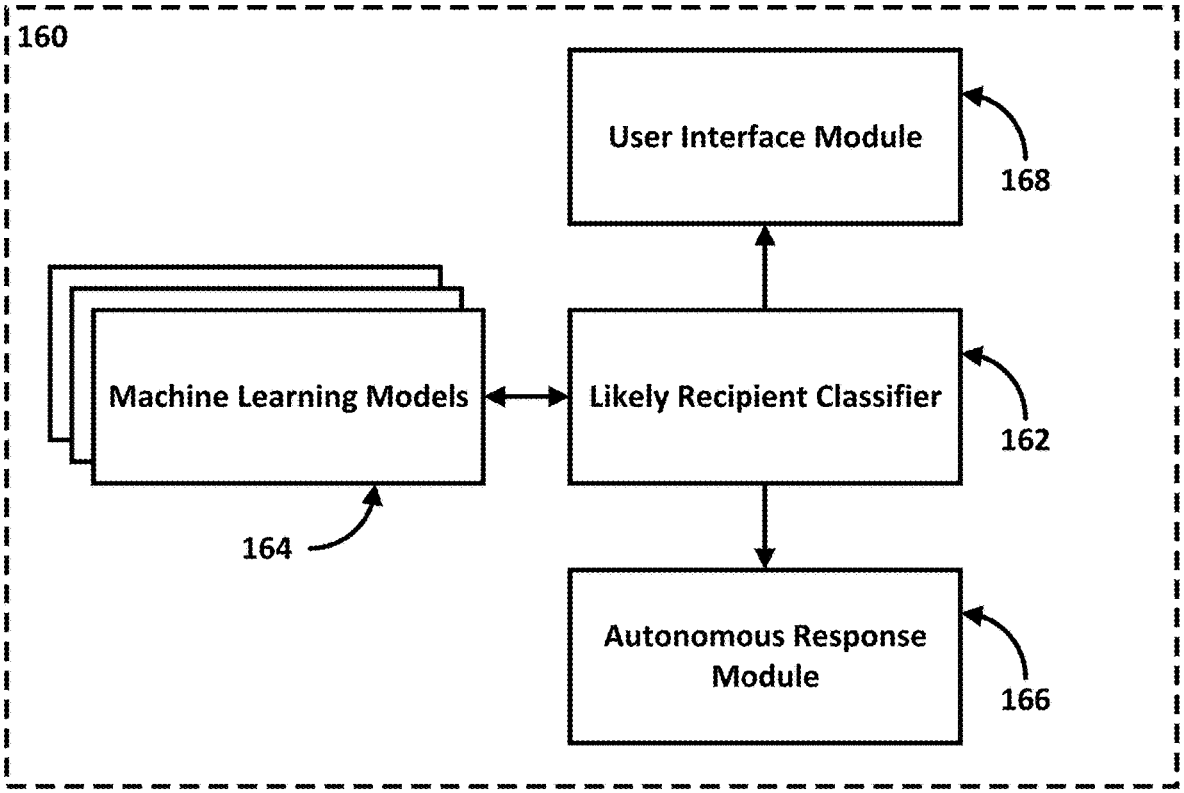


FIGURE 6

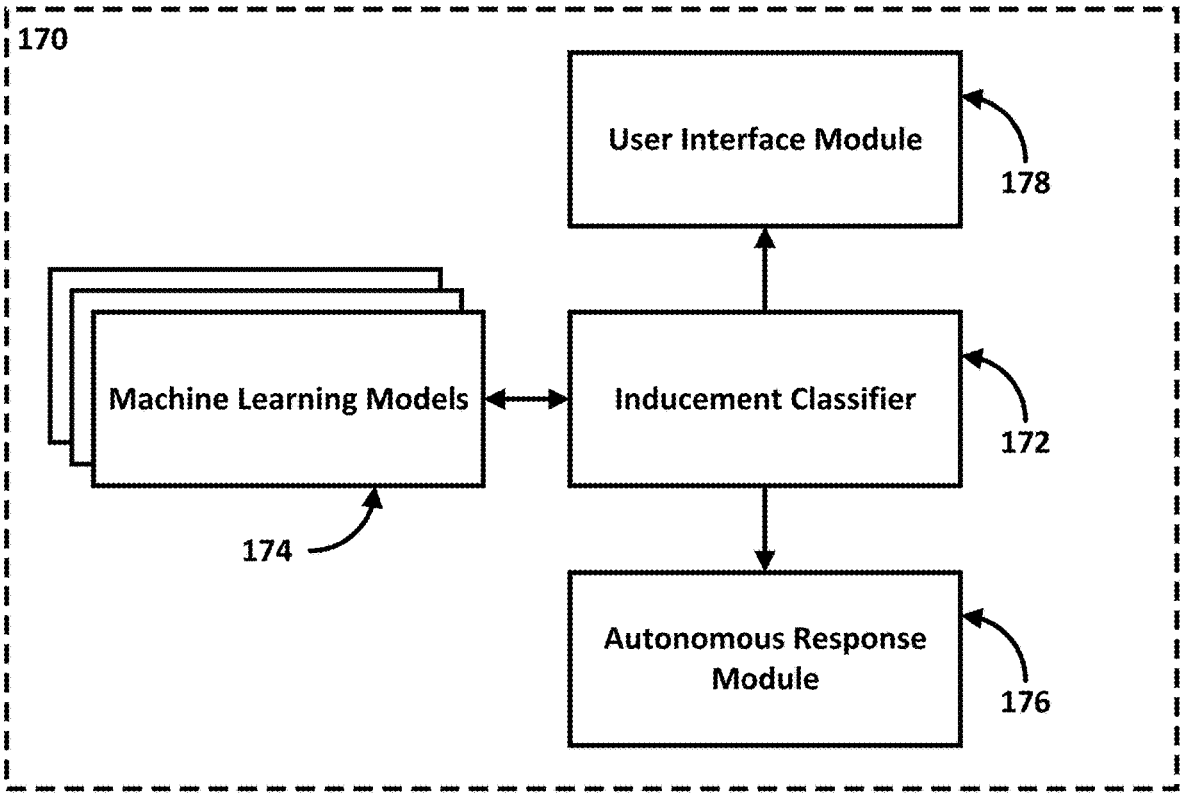


FIGURE 7A

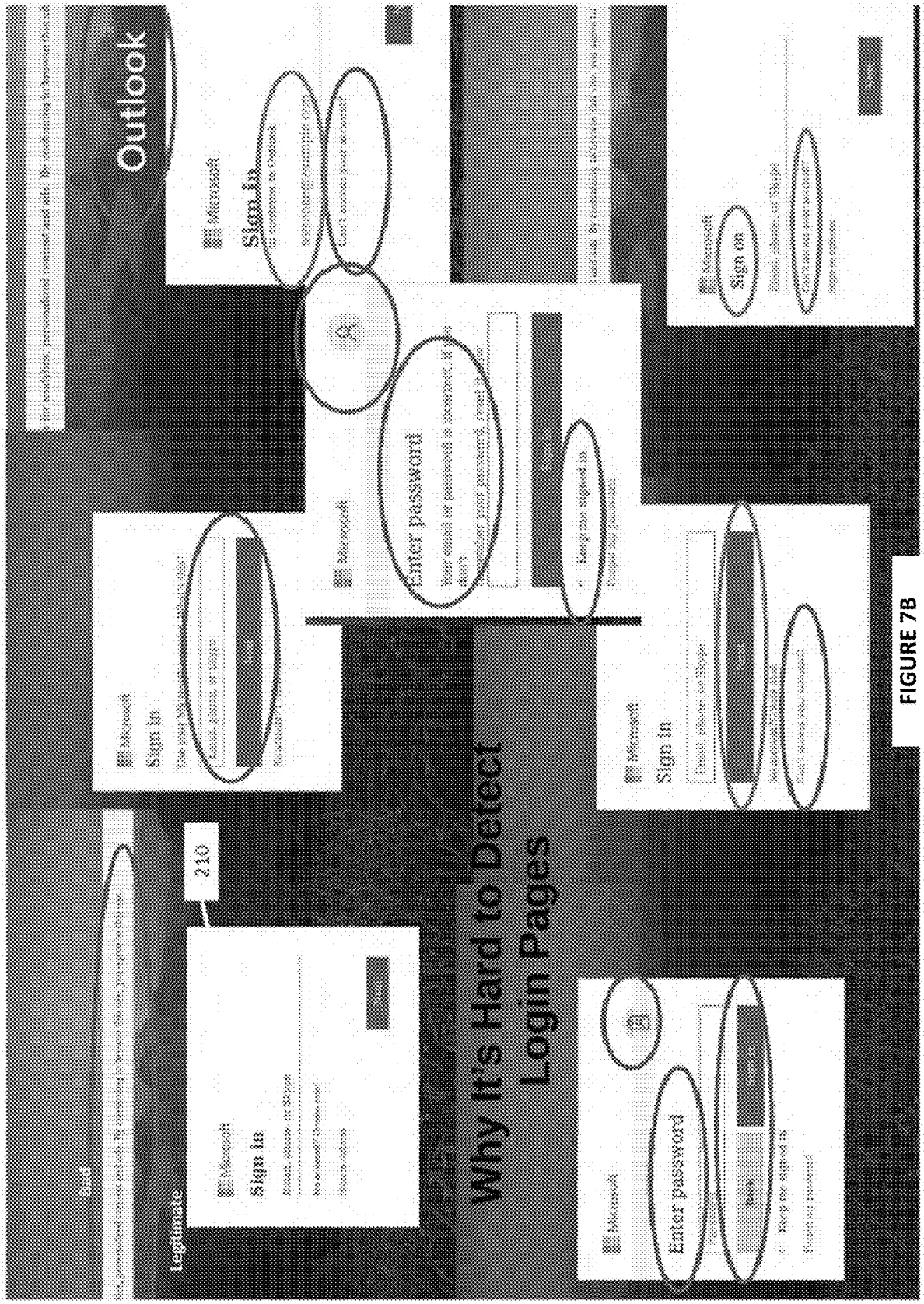

Why It's Hard to Detect Login Pages
Bad
Legitimate
210
Outlook
Microsoft
Sign in
Enter password
Sign on

FIGURE 7B

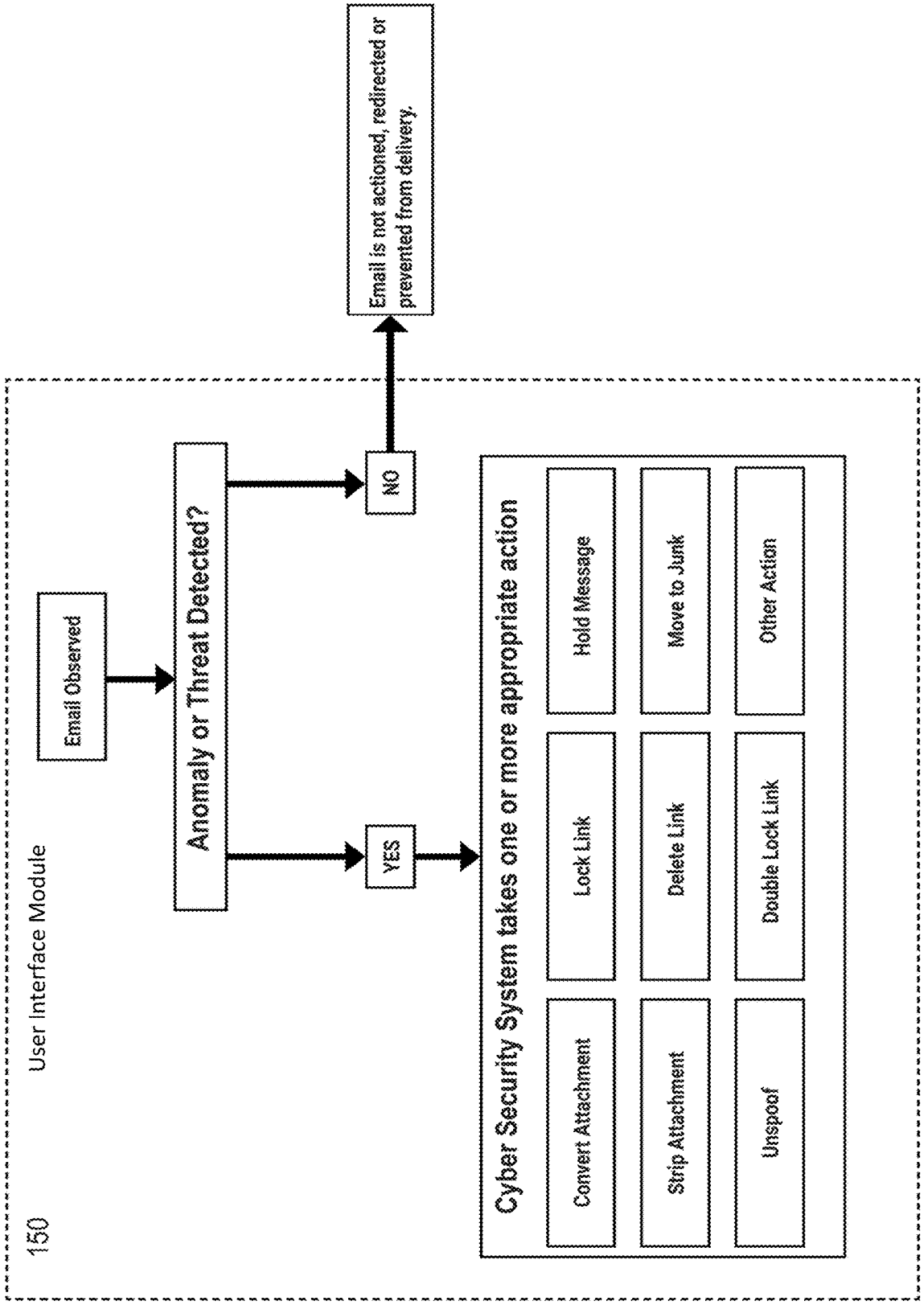

150
User Interface Module
Email Observed
Anomaly or Threat Detected?
NO
Email is not actioned, redirected or prevented from delivery.
YES
Cyber Security System takes one or more appropriate action
Convert Attachment
Lock Link
Hold Message
Strip Attachment
Delete Link
Move to Junk
Unspoof
Double Lock Link
Other Action

FIGURE 8

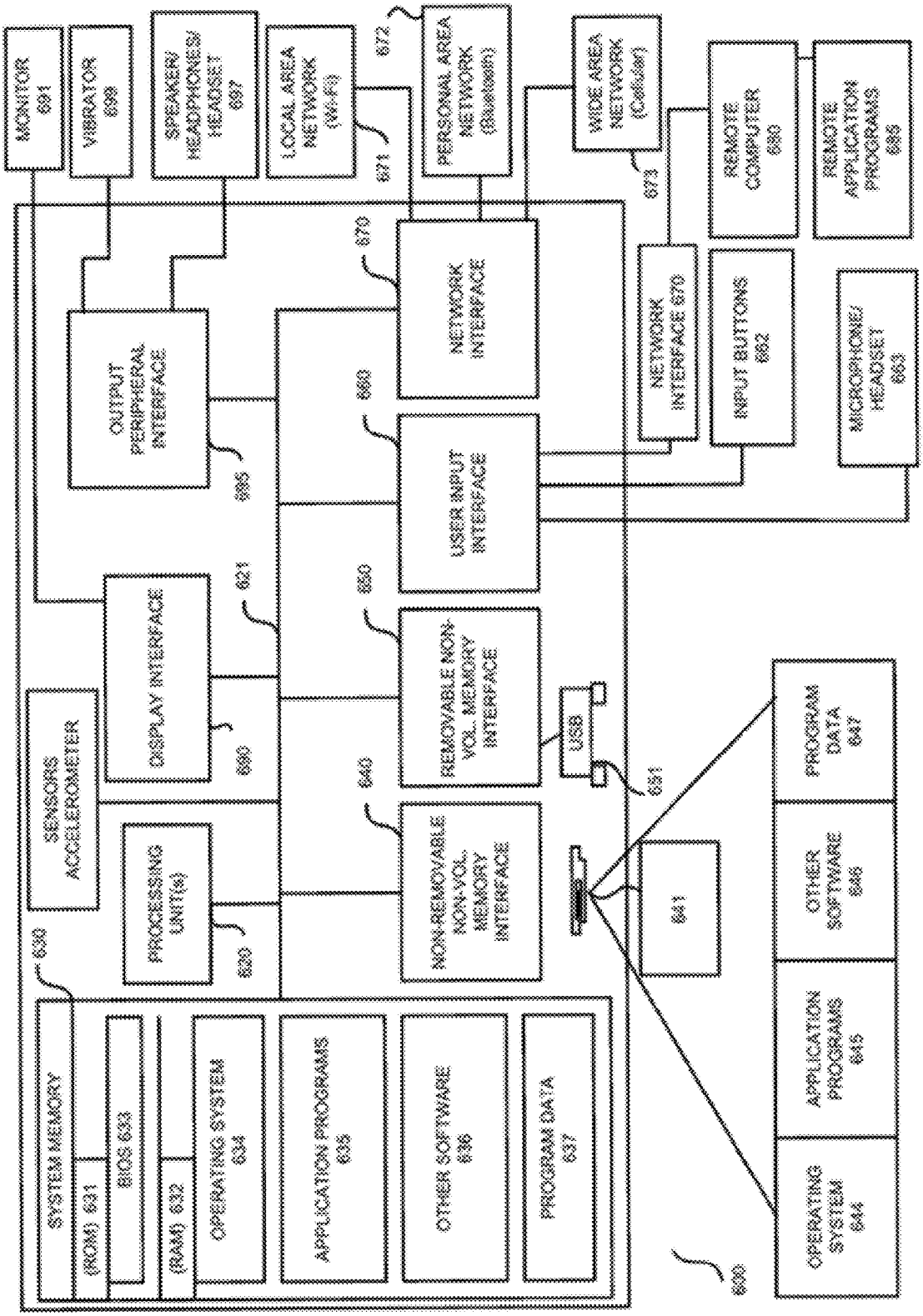

MONITOR 691
VIBRATOR 699
SPEAKER/ HEADPHONES/ HEADSET 697
LOCAL AREA NETWORK (WiFi)
671
672
PERSONAL AREA NETWORK (Bluetooth)
WIDE AREA NETWORK (Cellular)
673
REMOTE COMPUTER 680
REMOTE APPLICATION PROGRAMS 685
OUTPUT PERIPHERAL INTERFACE
695
NETWORK INTERFACE
670
660
NETWORK INTERFACE 670
INPUT BUTTONS 662
MICROPHONE/ HEADSET 663
USER INPUT INTERFACE
650
621
DISPLAY INTERFACE
690
SENSORS ACCELEROMETER
REMOVABLE NON-VOL. MEMORY INTERFACE
USB
691
640
PROCESSING UNIT(s)
620
NON-REMOVABLE NON-VOL. MEMORY INTERFACE
641
PROGRAM DATA 647
OTHER SOFTWARE 646
APPLICATION PROGRAMS 645
OPERATING SYSTEM 644
630
SYSTEM MEMORY
(ROM) 631
BIOS 633
(RAM) 632
OPERATING SYSTEM 634
APPLICATION PROGRAMS 635
OTHER SOFTWARE 636
PROGRAM DATA 637
600

FIGURE 9

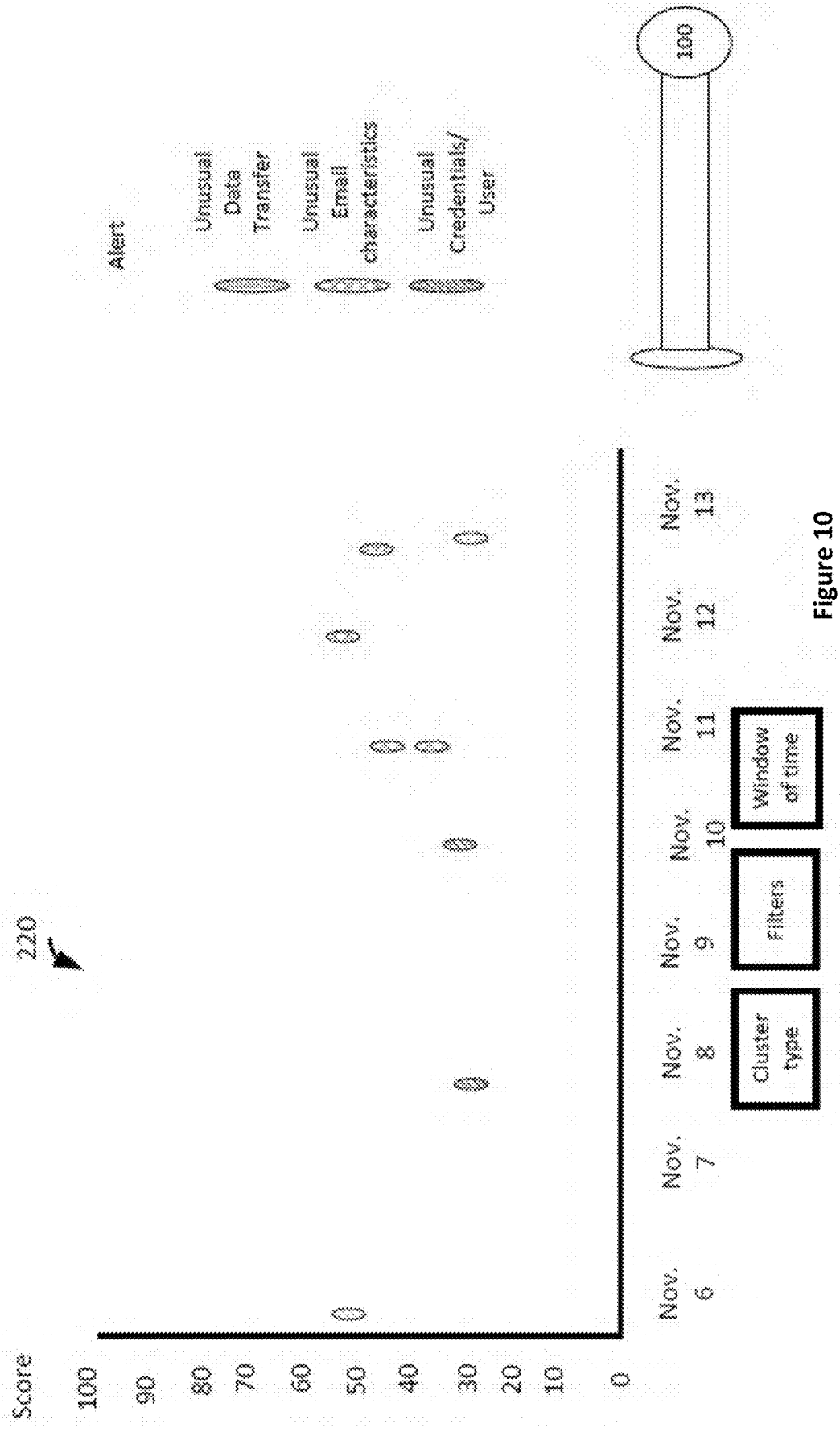

Score
100
90
80
70
60
50
40
30
20
10
0
220
Nov. 6
Nov. 7
Nov. 8
Nov. 9
Nov. 10
Nov. 11
Nov. 12
Nov. 13
Cluster type
Filters
Window of time
Alert
Unusual Data Transfer
Unusual Email characteristics
Unusual Credentials/ User
100

Figure 10

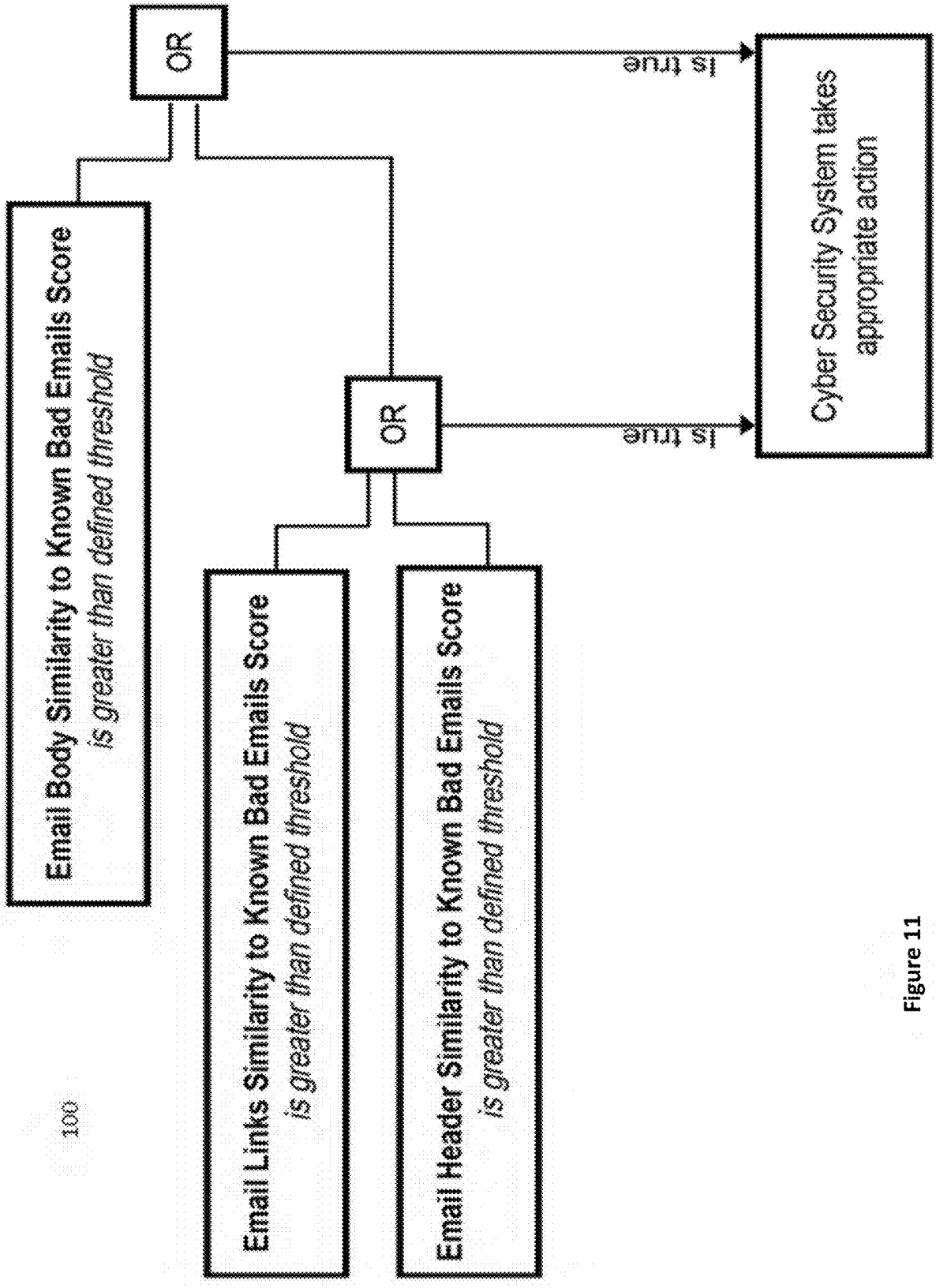

100
Email Body Similarity to Known Bad Emails Score
is greater than defined threshold
Email Links Similarity to Known Bad Emails Score
is greater than defined threshold
Email Header Similarity to Known Bad Emails Score
is greater than defined threshold
OR
OR
is true
is true
Cyber Security System takes appropriate action

Figure 11

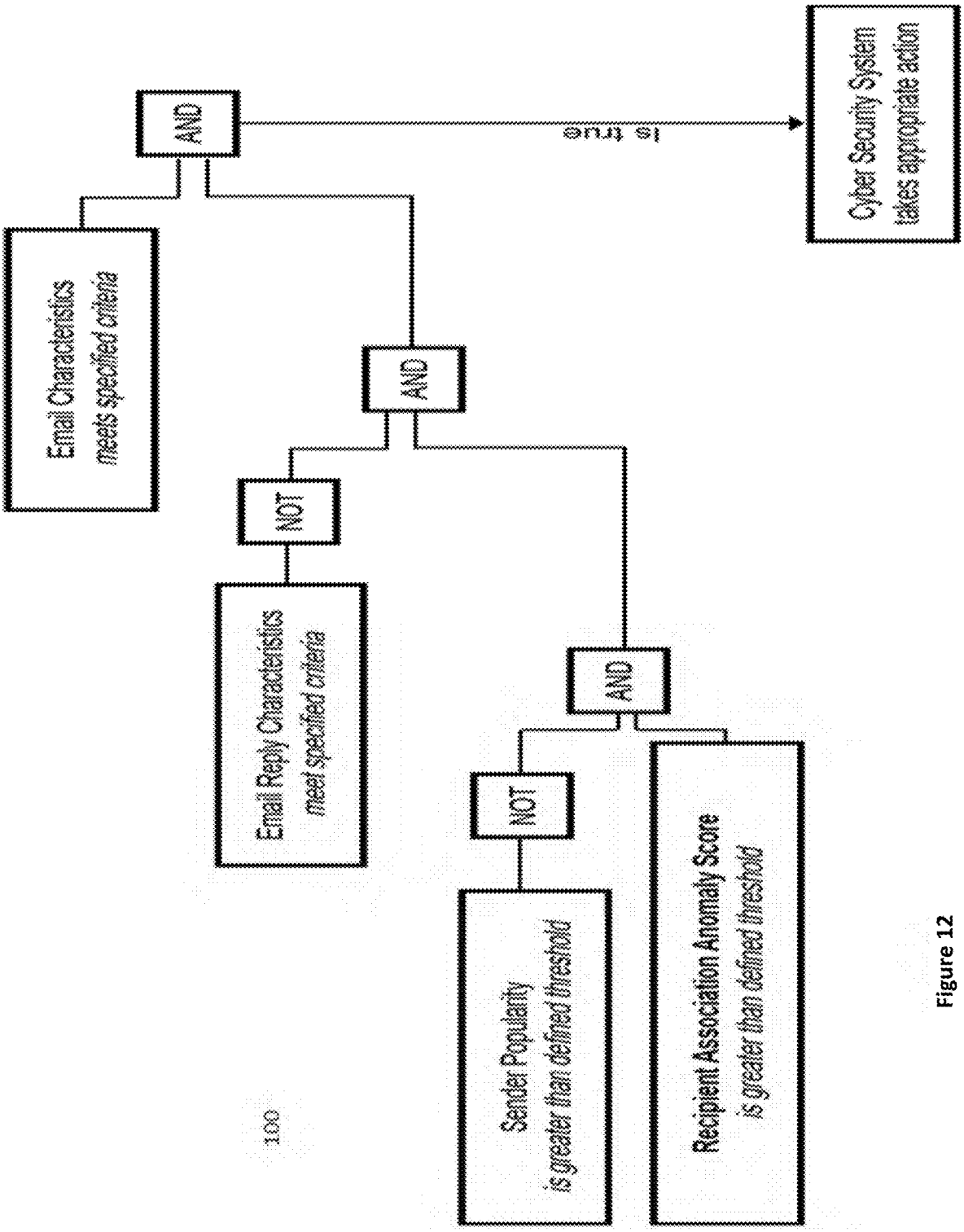
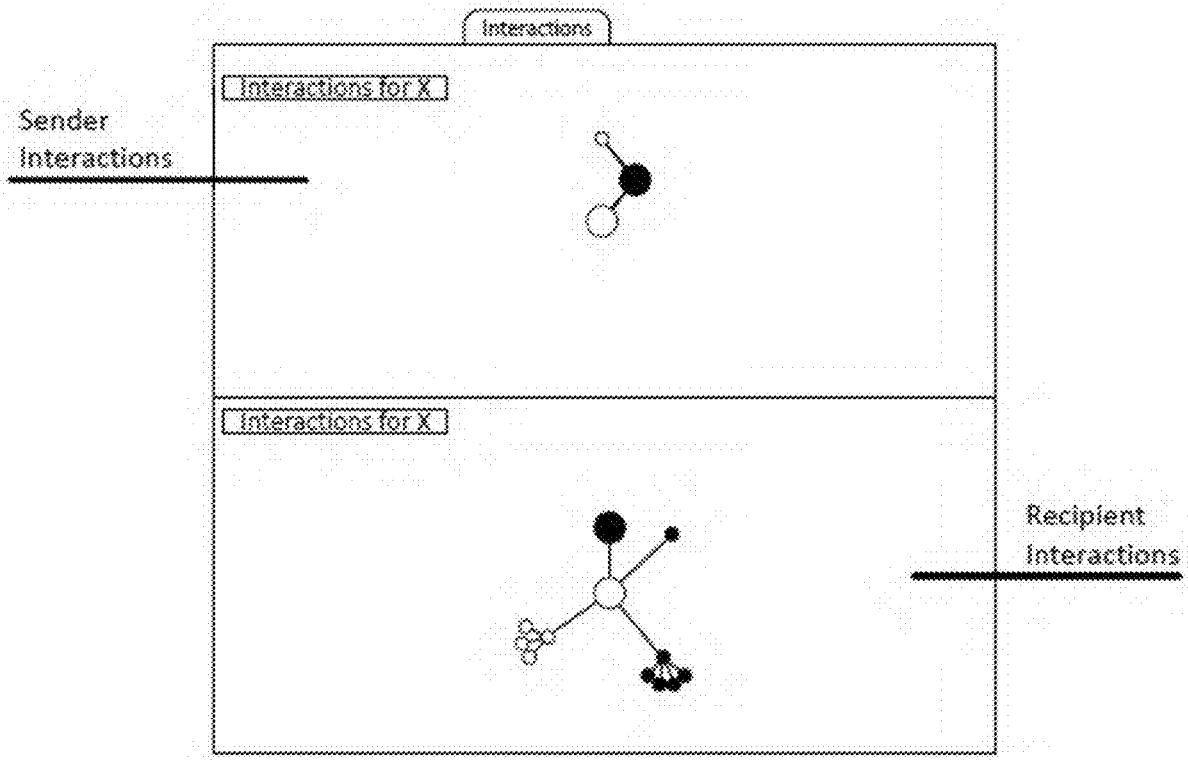

AND
is true
Cyber Security System takes appropriate action
Email Characteristics meets specified criteria
AND
NOT
Email Reply Characteristics meet specified criteria
AND
NOT
Sender Popularity is greater than defined threshold
Recipient Association Anomaly Score is greater than defined threshold
100

Figure 12


Interactions
Interactions for X
Sender
Interactions
Interactions for X
Recipient
Interactions

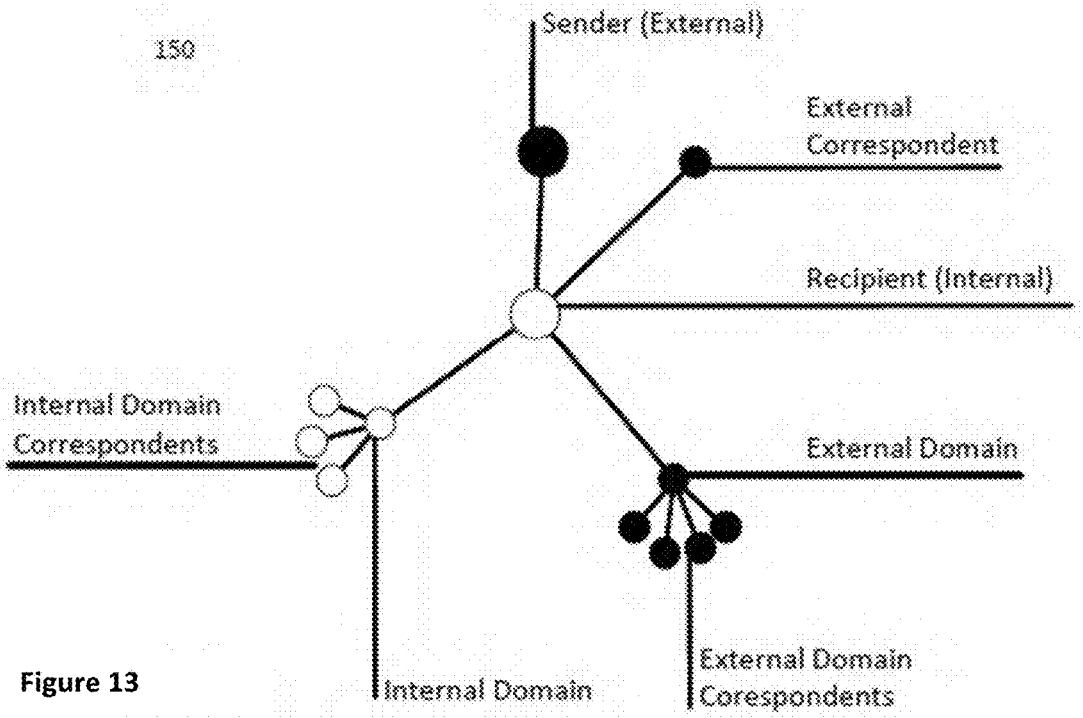

150
Sender (External)
External
Correspondent
Recipient (Internal)
Internal Domain
Correspondents
External Domain
Internal Domain
External Domain
Corespondents

Figure 13

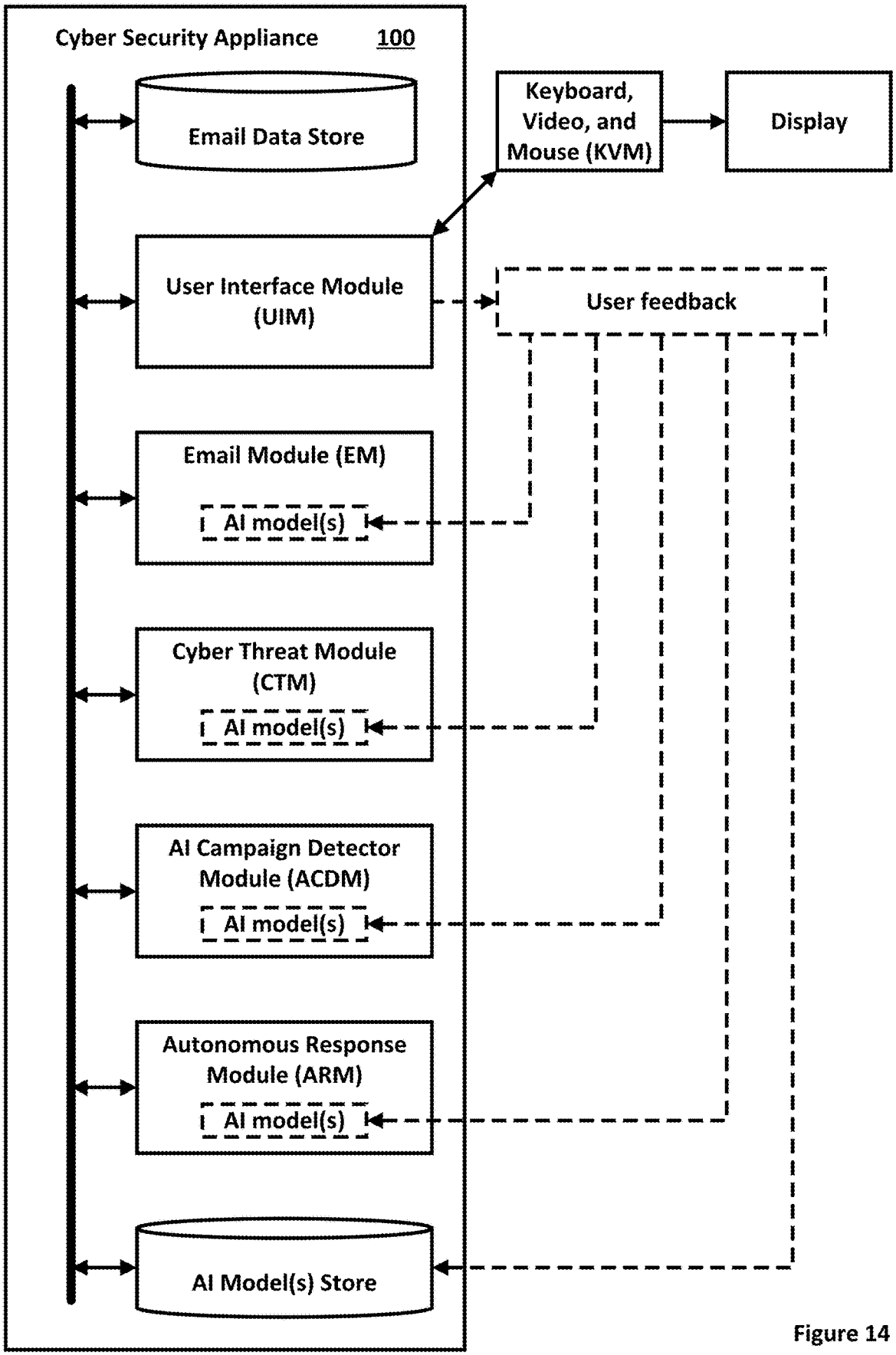

Cyber Security Appliance 100
Email Data Store
Keyboard, Video, and Mouse (KVM)
Display
User Interface Module (UIM)
User feedback
Email Module (EM)
AI model(s)
Cyber Threat Module (CTM)
AI model(s)
AI Campaign Detector Module (ACDM)
AI model(s)
Autonomous Response Module (ARM)
AI model(s)
AI Model(s) Store

Figure 14

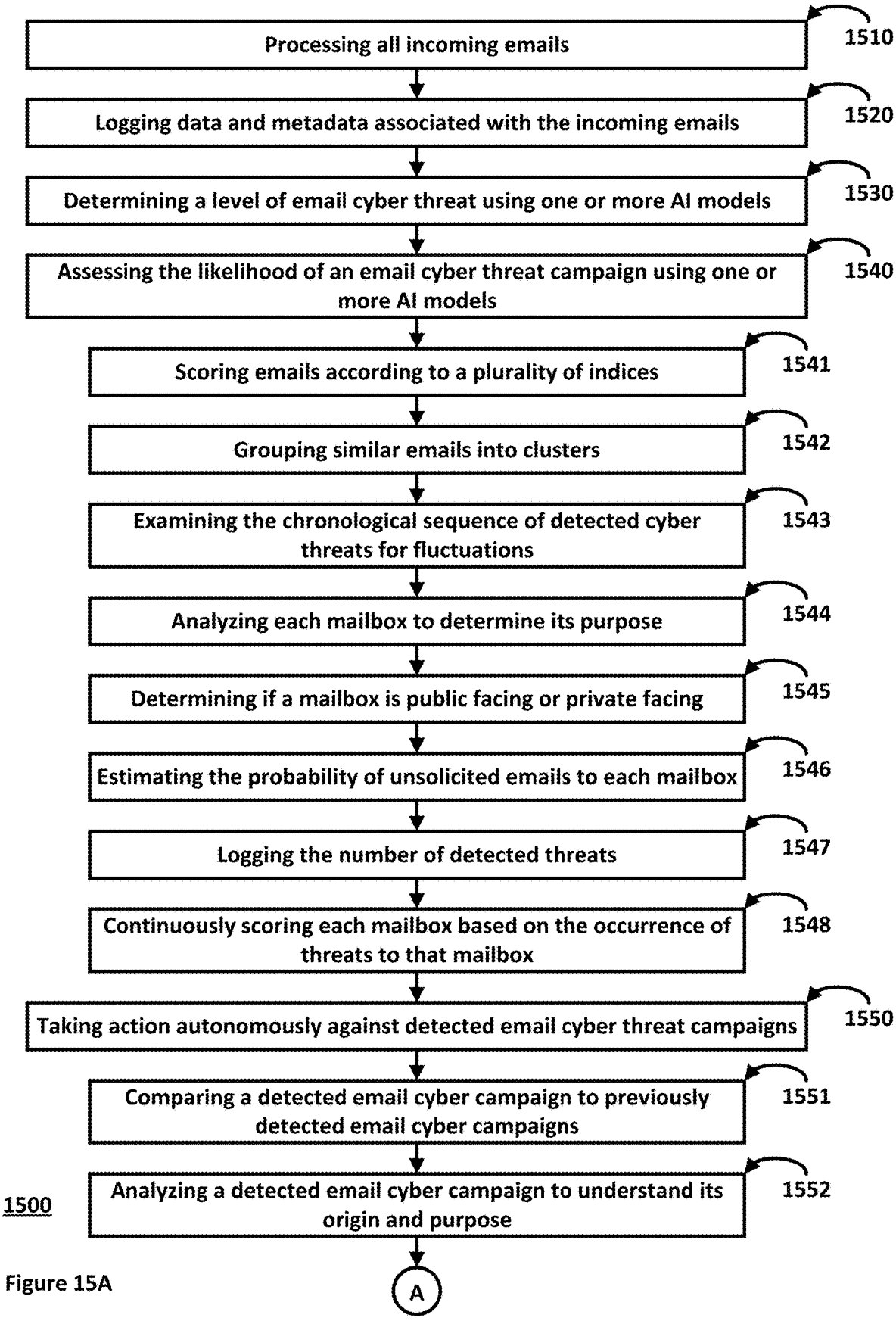

Processing all incoming emails
1510
Logging data and metadata associated with the incoming emails
1520
Determining a level of email cyber threat using one or more AI models
1530
Assessing the likelihood of an email cyber threat campaign using one or more AI models
1540
Scoring emails according to a plurality of indices
1541
Grouping similar emails into clusters
1542
Examining the chronological sequence of detected cyber threats for fluctuations
1543
Analyzing each mailbox to determine its purpose
1544
Determining if a mailbox is public facing or private facing
1545
Estimating the probability of unsolicited emails to each mailbox
1546
Logging the number of detected threats
1547
Continuously scoring each mailbox based on the occurrence of threats to that mailbox
1548
Taking action autonomously against detected email cyber threat campaigns
1550
Comparing a detected email cyber campaign to previously detected email cyber campaigns
1551
Analyzing a detected email cyber campaign to understand its origin and purpose
1552
1500
A

Figure 15A

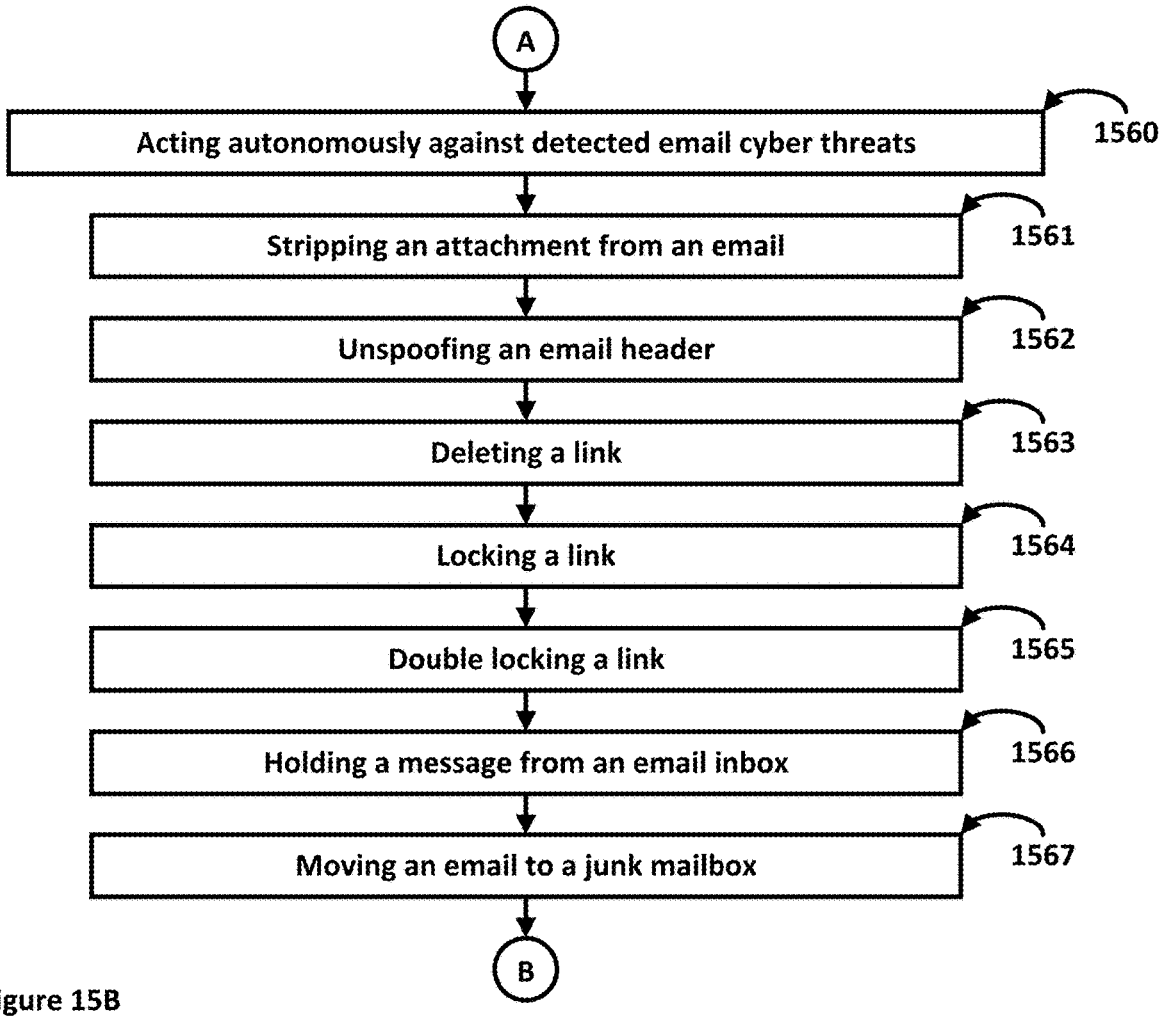

A
Acting autonomously against detected email cyber threats
1560
Stripping an attachment from an email
1561
Unspoofing an email header
1562
Deleting a link
1563
Locking a link
1564
Double locking a link
1565
Holding a message from an email inbox
1566
Moving an email to a junk mailbox
1567
B

Figure 15B

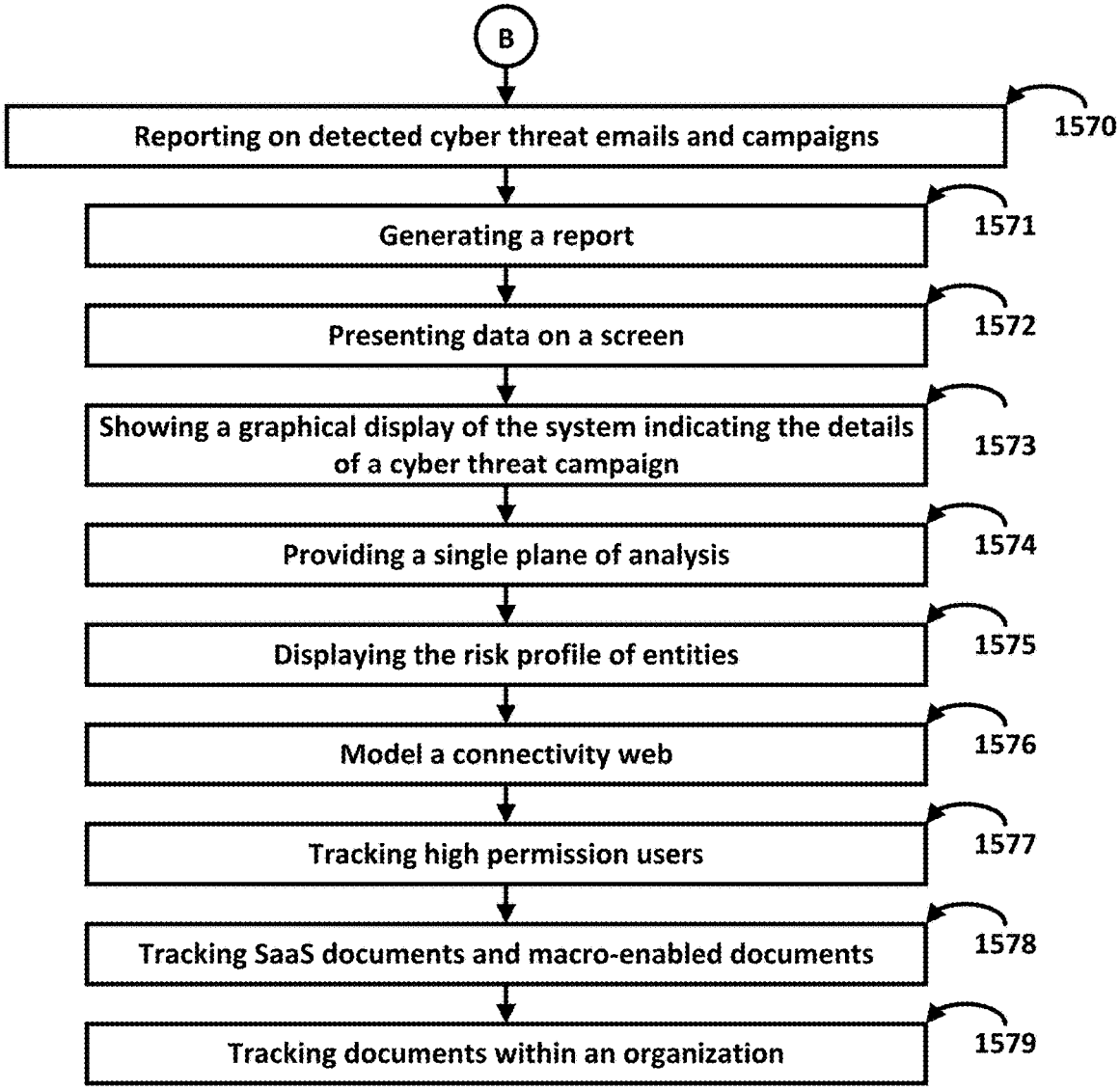

B
Reporting on detected cyber threat emails and campaigns
1570
Generating a report
1571
Presenting data on a screen
1572
Showing a graphical display of the system indicating the details of a cyber threat campaign
1573
Providing a single plane of analysis
1574
Displaying the risk profile of entities
1575
Model a connectivity web
1576
Tracking high permission users
1577
Tracking SaaS documents and macro-enabled documents
1578
Tracking documents within an organization
1579

Figure 15C

A METHOD AND SYSTEM FOR DETERMINING AND ACTING ON AN EMAIL CYBER THREAT CAMPAIGN

NOTICE OF COPYRIGHT

CROSS-REFERENCE TO RELATED APPLICATIONS

[0002] This application claims the benefit of and priority under 35 USC 119 to:

[0003] (A) U.S. Provisional Application No. 63/219,026 filed Jul. 7, 2021, entitled A CYBER SECURITY APPLIANCE AND OTHER SECURITY TOOLS;

[0004] (B) U.S. Provisional Application No. 63/317,157 filed Mar. 7, 2022, entitled A CYBER SECURITY SYSTEM;

[0005] (C) and priority under 35 USC 120 to U.S. Non-provisional application Ser. No. 17/187,381 filed Feb. 26, 2021, entitled A METHOD AND SYSTEM FOR DETERMINING AND ACTING ON A STRUCTURED DOCUMENT CYBER THREAT RISK, claiming priority to:

[0006] (i) U.S. Provisional Application No. 62/983,307 filed Feb. 28, 2020, entitled AN ARTIFICIAL INTELLIGENCE BASED CYBER SECURITY SYSTEM, and

[0007] (ii) U.S. Provisional Application No. 63/026,446 filed May 18, 2020, entitled A CYBER SECURITY SYSTEM USING ARTIFICIAL INTELLIGENCE; and

[0008] (D) U.S. Non-provisional application Ser. No. 16/732,644 filed Jan. 2, 2020, entitled A CYBER THREAT DEFENSE SYSTEM PROTECTING EMAIL NETWORKS WITH MACHINE LEARNING MODELS USING A RANGE OF METADATA FROM OBSERVED EMAIL COMMUNICATIONS, claiming priority to:

[0009] (i) U.S. Provisional Application No. 62/796,507 filed Jan. 24, 2019, entitled A CYBER SECURITY SYSTEM, claiming priority as a continuation-in-part to:

[0010] (ii) U.S. Non-provisional application Ser. No. 16/278,932 filed Feb. 19, 2019, entitled A CYBER THREAT DEFENSE SYSTEM PROTECTING EMAIL NETWORKS WITH MACHINE LEARNING MODELS, which claims priority to:

[0011] (iii) U.S. Provisional Application No. 62/632,623 filed Feb. 20, 2018, entitled A CYBER THREAT DEFENSE SYSTEM WITH VARIOUS IMPROVEMENTS;

[0012] the disclosure of each of which is hereby expressly incorporated by reference herein in its entirety.

FIELD

[0013] Embodiments of the design provided herein generally relate to a cyber threat defense system. In an embodiment, Artificial Intelligence is applied to analyzing Cyber Security threats coming from and/or associated with a structured document, such as an email, an instant message, a text message, or other structured electronic communication and campaigns related to these threats.

BACKGROUND

[0014] In the cyber security environment, firewalls, endpoint security methods, and other tools such as SIEMs and sandboxes are deployed to enforce specific policies and provide protection against certain threats. These tools currently form an important part of an organization's cyber defense strategy, but they are insufficient in the new age of cyber threats. Legacy tools are failing to deal with new cyber threats because the traditional approach relies on being able to pre-define the cyber threat in advance by writing rules or producing signatures. In today's environment, this approach to defending against cyber threats is fundamentally flawed:

[0015] Threats are constantly evolving—novel attacks do not match historical-attack "signatures," and even subtle changes to previously understood attacks can result in them going undetected by legacy defenses;

[0016] Rules and policies defined by organizations are continually insufficient—security teams simply cannot imagine every possible thing that may go wrong in the future; and

[0017] Employee 'insider' threat is a growing trend—it is difficult to spot malicious employees behaving inappropriately as they are a legitimate presence on the business network.

[0018] The reality is that modern threats bypass the traditional legacy defense tools on a daily basis. These tools need a new tool based on a new approach that can complement them and mitigate their deficiencies at scale across the entirety of digital organizations. In the complex modern world, it is advantageous that the approach is fully automated as it is virtually impossible for humans to sift through the vast amount of security information gathered each minute within a digital.

SUMMARY

[0019] In an embodiment, various methods, apparatuses, and systems are discussed for a cyber security system to protect from cyber threat risks in relation to outgoing or incoming structured documents that are addressed to a recipient by a sender and campaigns related to these threats. One or more machine learning models are trained on the classification of structured documents with one or more of a plurality of categories based on a plurality of characteristics of the structured documents. A classifier is configured to receive a structured document for analysis and to parse the structured document to extract the plurality of characteristics of the structured document. The classifier is further configured to classify the structured document with one or more of the plurality of categories based on the extracted plurality of characteristics and the one or more machine learning models and to determine an associated score for the classification. An autonomous response module is then configured to, based on a comparison of the associated score with a threshold, cause one or more autonomous actions to be taken in relation to the structured document.

[0020] In one aspect of the present disclosure, the structured document is to be sent from the sender to an indicated recipient; each category of the plurality of categories represents a respective recipient of a plurality of recipients known to the sender; and the associated score represents the probability of a match between the indicated recipient and the extracted plurality of characteristics. In this aspect, the classifier may be configured to determine one or more

further scores representing the respective probability of a match between the extracted plurality of characteristics and each of the other recipients known to the sender; and the threshold may represent the score of an alternative recipient, of the other recipients known to the sender, having the highest probability of a match. Moreover, the one or more autonomous actions may comprise, if the associated score is less than the threshold, displaying an alert to the sender on the sender user interface indicating that the alternative recipient has a higher probability of a match than the indicated recipient.

[0021] In this manner, the system of this aspect may use the one or more machine learning models to determine a set of scores associated with the respective matches between the characteristics extracted from the structured document and characteristic signatures associated with the recipient indicated in a structured document to be sent as well as other recipients known to the sending user. Where a recipient other than the indicated recipient is determined to have the best match/highest score, in respect of the extracted characteristics, this may be brought to the attention of the sender by displaying an alert on a sender user interface.

[0022] In a particular example embodiment of this aspect, the one or more of the machine learning models are trained to identify, for each recipient known to the sender, one or more indicators corresponding to characteristics that are frequently present in structured documents sent by the sender and addressed to the respective recipient known to the sender relative to those addressed to other recipients known to the sender; and the classifier classifies the structured document with one or more of the categories representing the plurality of recipients known to the sender by comparing the extracted plurality of characteristics with the one or more indicators for each recipient known to the sender.

[0023] In a further aspect of the present disclosure, the structured document has instead been sent to a user from a given sender. In this aspect, the one or more categories may comprise one or more malign categories and, when the associated score determined for the one or more malign categories is above the threshold, the one or more autonomous actions comprise one or more actions to contain the malign nature of the sent structured document.

[0024] In some embodiments of the present disclosure, an AI classifier is used to analyze patterns of emails and other structured documents deemed malign by the system in order to detect a cyber threat email/structure document campaign comprising multiple emails/structured documents. An analysis is performed to understand the nature and purpose of such a campaign, identify the bad actors responsible for the campaign, and take appropriate action in unison with the autonomous response module.

[0025] In this manner, the system of this further aspect of the present disclosure may determine whether an incoming structured document represents a potential cyber threat and what type of cyber threat this may be. Appropriate autonomous actions may then be performed by the system to contain or neutralize the potential cyber threat associated with the incoming structured document. By pulling together a wide variety of characteristics, the apparatus and system can provide an improved granularity in the classification of these structured documents, and this then results in the ability to provide improved decision making/autonomous actions in response to this classification.

[0026] These and other features of the design provided herein can be better understood with reference to the drawings, description, and claims, all of which form the disclosure of this patent application.

DRAWINGS

[0027] The drawings refer to some embodiments of the design provided herein in. In particular:

[0028] FIG. **1** illustrates a block diagram of an example cyber security appliance;

[0029] FIG. **2** illustrates a block diagram of an example cyber security appliance monitoring email activity and network activity to feed this data to correlate causal links between these activities to supply this input into the cyber threat analysis;

[0030] FIG. **3** illustrates a block diagram of an example of the cyber threat module determining a threat risk parameter that factors in how the chain of unusual behaviors correlates to potential cyber threats and 'the likelihood that this chain of one or more unusual behaviors of the email activity and user activity under analysis fall outside of derived normal benign behavior;' and thus, is malicious behavior;

[0031] FIG. **4** illustrates a block diagram of an example of the cyber threat defense system referencing one or more machine learning models trained on gaining an understanding of a plurality of characteristics of an email itself and its related data, including classifying the properties of the email and its metadata;

[0032] FIG. **5** illustrates an example cyber threat defense system protecting an example network of computer systems;

[0033] FIG. **6** illustrates an apparatus according to an aspect of the present disclosure;

[0034] FIG. **7**A illustrates an apparatus according to another aspect of the present disclosure;

[0035] FIG. **7**B illustrates a diagram of an embodiment of a segmentation module employed by a phishing site detector to analyze key text-like features of an example legitimate site and multiple example legitimate sites;

[0036] FIG. **8** illustrates a block diagram of an embodiment of example autonomous actions that the autonomous response module can be configured to take without a human initiating that action;

[0037] FIG. **9** illustrates a block diagram of an embodiment of one or more computing devices that can be used in combination with the present disclosure;

[0038] FIG. **10** illustrates a block diagram of an embodiment of an example chain of unusual behavior for the email(s) in connection with the rest of the network under analysis;

[0039] FIG. **11** illustrates a block diagram of an embodiment of an email similarity scoring module configured to cooperate with the one or more mathematical models in order to compare an incoming email based on semantic similarity of multiple aspects of the email to a cluster of different metrics derived from known bad emails to derive a similarity score between an email under analysis and the cluster of different metrics derived from known bad emails;

[0040] FIG. **12** illustrates a block diagram of an embodiment of a mass email association detector configured to determine a similarity between two or more highly similar emails being i) sent from or ii) received by a collection of two or more individual users in the email domain in a substantially simultaneous time frame, where one or more

mathematical models are used to determine similarity weighing in order to derive a similarity score between compared emails;

[0041] FIG. **13** illustrates a block diagram of an embodiment of an email module and network module cooperating to supply a particular user's network activity tied to their email activity;

[0042] FIG. **14** illustrates a block diagram of an example cyber security appliance configured to detect, analyze, and act against email cyber threat campaigns; and

[0043] FIGS. **15**A, **15**B, and **15**C illustrate a flowchart diagram depicting a process of the operation of cyber security appliance configured to detect, analyze, and act against email cyber threat campaigns.

[0044] While the design is subject to various modifications, equivalents, and alternative forms, specific embodiments thereof have been shown by way of example in the drawings and will now be described in detail. It should be understood that the design is not limited to the particular embodiments disclosed, but—on the contrary—the intention is to cover all modifications, equivalents, and alternative forms using the specific embodiments.

DESCRIPTION

[0045] In the following description, numerous specific details are set forth, such as examples of specific data signals, named components, number of servers in a system, etc., in order to provide a thorough understanding of the present design. It will be apparent, however, to one of ordinary skill in the art that the present design can be practiced without these specific details. In other instances, well-known components or methods have not been described in detail but rather in a block diagram in order to avoid unnecessarily obscuring the present design. Further, specific numeric references such as a first server can be made. However, the specific numeric reference should not be interpreted as a literal sequential order but rather interpreted that the first server is different than a second server. Thus, the specific details set forth are merely exemplary. Also, the features implemented in one embodiment may be implemented in another embodiment where logically possible. The specific details can be varied from and still be contemplated to be within the spirit and scope of the present design. The term coupled is defined as meaning connected either directly to the component or indirectly to the component through another component.

[0046] In general, artificial intelligence is used to analyze cyber security threats in the present disclosure. A cyber defense system can use models that are trained on a wide range of characteristics extracted from structured documents, such as an email, an instant message, a text message, or other structured electronic communication. The following disclosure will describe a cyber defense system implemented in relation to emails; however, it will be appreciated by the skilled person that this teaching could easily be translated to other types of structured documents without departing from the scope of the present application.

[0047] Further models of the cyber defense system may be trained on the normal behavior of email activity and user activity associated with an email system. A cyber threat module may reference the models that are trained on the normal behavior of email activity and user activity. A determination is made of a threat risk parameter that factors in the likelihood that a chain of one or more unusual behaviors of the email activity and user activity under analysis fall outside of derived normal benign behavior. An autonomous response module can be used, rather than a human taking action, to cause one or more autonomous rapid actions to be taken to contain the cyber threat when the threat risk parameter from the cyber threat module is equal to or above an actionable threshold.

Example Cyber Security Appliance

[0048] FIG. **1** illustrates a selection of modules forming at least part of an example cyber security appliance/cyber threat system. Various Artificial Intelligence models and modules of the cyber security appliance/cyber threat system cooperate to protect a system, including but not limited to an email network, from cyber threats. The Cyber Security Appliance (CSA) **100** may include a trigger module, a gatherer module, an analyzer module, an assessment module, a formatting module, an autonomous report composer, a data store, one or more Artificial Intelligence models trained on potential cyber threats and their characteristics, symptoms, remediations, etc., one or more Artificial Intelligence models trained with machine learning on a normal pattern of life for entities in the network, one or more Artificial Intelligence models trained with machine learning on threat report generation, and multiple libraries of text and visual representations to cooperate the library of page templates to populate visual representations, such as graphs, and text on the pages of the threat report.

[0049] The one or more machine learning models may include a first Artificial Intelligence model trained on characteristics of an email itself and its related data, a second Artificial Intelligence model trained on potential cyber threats, and one or more Artificial Intelligence models, each trained on different users, devices, system activities and interactions between entities in the system, other aspects of the system, and the normal "pattern of life" in the system. An example network of an email system will be used to illustrate portions of a CSA **100**. However, it will be appreciated by the skilled person that this teaching could easily be translated to systems for handling types of structured documents other than emails without departing from the scope of the present application—for example, an instant message, a text message, or other structured electronic communication

[0050] Referring to FIG. **1**, the trigger module may detect time-stamped data indicating an event is occurring, and it may then be triggered that something unusual is happening. The gatherer module is triggered by specific events or alerts of i) an abnormal behavior, ii) a suspicious activity, and iii) any combination of both. The trigger module may identify, with one or more AI models trained with machine learning on a normal email pattern of life for entities in the email network, at least one of i) an abnormal behavior, ii) a suspicious activity, and iii) any combination of both, from one or more entities in the system.

[0051] The inline data may be gathered on the deployment from a data store when the traffic is observed. The gatherer module may initiate a collection of data to support or refute each of the one or more possible cyber threat hypotheses that could include this abnormal behavior or suspicious activity by the one or more AI models trained on possible cyber threats. The gatherer module cooperates with a data store. The data store stores comprehensive logs for network traffic observed. These logs can be filtered with complex logical

queries, and each IP packet can be interrogated on a vast number of metrics in the network information stored in the data store.

[0052] The data store can store the metrics and previous threat alerts associated with network traffic for a period of time (for example, 27 days may be set as a default value in some embodiments). This corpus of data is fully searchable. The CSA **100** works with network probes to monitor network traffic and store and record the data and metadata associated with the network traffic in the data store.

[0053] FIG. **2** illustrates an example of CSA using an intelligent-adversary simulator cooperating with a network module and network probes ingesting traffic data for network devices and network users in the network under analysis.

[0054] Referring back to FIG. **1**, the gatherer module may consist of multiple automatic data gatherers that each look at different aspects of the data depending on the particular hypothesis formed for the analyzed event. The data relevant to each type of possible hypothesis can be automatically pulled from additional external and internal sources. Some data is pulled or retrieved by the gatherer module for each possible hypothesis.

[0055] The gatherer module may further extract data, at the request of the analyzer module, on each possible hypothetical threat that would include the abnormal behavior or suspicious activity; and then filter that collection of data down to relevant points of data to either 1) support or 2) refute each particular hypothesis of what the potential cyber threat, e.g., the suspicious activity and/or abnormal behavior, relates to. The gatherer module and the data store can cooperate to store an inbound and outbound email flow received over a period of time as well as autonomous actions performed by the autonomous response module on that email flow. The gatherer module may send the filtered down relevant points of data to either 1) support or 2) refute each particular hypothesis to the analyzer module, comprised of one or more algorithms used by the AI models trained with machine learning on possible cyber threats to make a determination on a probable likelihood of whether that particular hypothesis is supported or refuted.

[0056] A feedback loop of cooperation between the gatherer module and the analyzer module may be used to apply one or more models trained on different aspects of this process.

[0057] The analyzer module can form one or more hypotheses on what are a possible set of activities, including cyber threats that could include the identified abnormal behavior and/or suspicious activity from the trigger module with one or more AI models trained with machine learning on possible cyber threats. The analyzer module may request further data from the gatherer module to perform this analysis. The analyzer module can cooperate with the one or more Artificial Intelligence models trained with machine learning on the normal email pattern of life for entities in the email network to detect anomalous email which is detected as outside the usual pattern of life for each entity, such as a user, of the email network. The analyzer module can cooperate with the Artificial Intelligence models trained on potential cyber threats to detect suspicious emails that exhibit traits that may suggest malicious intent, such as phishing links, scam language, sent from suspicious domains, etc. In addition, the gatherer module and the analyzer module may use a set of scripts to extract data on each possible hypothetical threat to supply to the analyzer module. The gatherer module and analyzer module may use a plurality of scripts to walk through a step-by-step process of what to collect to filter down to the relevant data points (from the potentially millions of data points occurring in the network) to make a decision on what is required by the analyzer module.

[0058] The analyzer module may further analyze a collection of system data, including metrics data, to support or refute each of the one or more possible cyber threat hypotheses that could include the identified abnormal behavior and/or suspicious activity data with the one or more AI models trained with machine learning on possible cyber threats. The analyzer module then generates at least one or more supported possible cyber threat hypotheses from the possible set of cyber threat hypotheses, and as well as could include some hypotheses that were not supported/refuted.

[0059] The analyzer module may get threat information from Open Source APIs as well as from databases as well as information trained into AI models. Also, probes collect the user activity and the email activity and then feed that activity to the network module to draw an understanding of the email activity and user activity in the email system.

[0060] The analyzer module learns how expert humans tackle investigations into specific cyber threats. The analyzer module may use i) one or more AI models and/or ii) rules-based models, and iii) combinations of both that are hosted within the plug-in appliance connecting to the network.

[0061] The AI models use data sources, such as simulations, database records, and actual monitoring of different human exemplar cases, as input to train the AI model on how to make a decision. The analyzer module also may utilize repetitive feedback, as time goes on, for the AI models trained with machine learning on possible cyber threats via reviewing a subsequent resulting analysis of the supported possible cyber threat hypothesis and supplying that information to the training of the AI models trained with machine learning on possible cyber threats in order to reinforce the model's finding as correct or inaccurate.

[0062] Each hypothesis of typical threats, e.g., human user insider attack/inappropriate network and/or email behavior, malicious software/malware attack/inappropriate network and/or email behavior, can have various supporting points of data and other metrics associated with that possible threat, and a machine learning algorithm will look at the relevant points of data to support or refute that particular hypothesis of what the suspicious activity and/or abnormal behavior relates to. Networks have a wealth of data and metrics that can be collected, and then the mass of data is filtered/condensed down into the important features/salient features of data by the gatherers.

[0063] The analyzer module may perform an analysis of internal and external data, including readout from machine learning models, which output a likelihood of the suspicious activity and/or abnormal behavior related for each hypothesis on what the suspicious activity and/or abnormal behavior relates to with other supporting data to support or refute that hypothesis.

[0064] The assessment module may assign a probability, or confidence level/associated score, of a given cyber threat hypothesis that is supported, and a threat level posed by that cyber threat hypothesis, which includes this abnormal behavior or suspicious activity, with the one or more AI models trained on possible cyber threats. The assessment

module can cooperate with the autonomous response module to determine an appropriate response to mitigate various cyber-attacks that could be occurring.

[0065] The analyzer module can reference machine learning models that are trained on the normal behavior of email activity and user activity associated with at least the email system, where the analyzer module cooperates with the assessment module to determine a threat risk parameter that factors in 'the likelihood that a chain of one or more unusual behaviors of the email activity and user activity under analysis fall outside of derived normal, benign behavior;' and thus, are likely malicious behavior.

[0066] For example, the one or more machine learning models can be self-learning models using unsupervised learning and trained on the normal behavior of different aspects of the system, for example, email activity and user activity associated with an email system. The self-learning models of normal behavior are regularly updated. The self-learning model of normal behavior is updated when new input data is received that is deemed within the limits of normal behavior. A normal behavior threshold is used by the model as a moving benchmark of parameters that correspond to a normal pattern of life for the computing system. The normal behavior threshold is varied according to the updated changes in the computer system, allowing the model to spot behavior on the computing system that falls outside the parameters set by the moving benchmark.

[0067] The cyber security appliance/cyber threat system is configurable in a user interface by a user, enabling what type of automatic response actions, if any, the cyber security appliance may take when different types of cyber threats, indicated by the pattern of behaviors under analysis, which are equal to or above a configurable level of threat posed by this malicious actor.

[0068] The cyber threat defense system **100** may also include one or more machine learning models trained on gaining an understanding of a plurality of characteristics of an email itself and its related data, including classifying the properties of the email and its metadata.

[0069] The cyber threat module can also reference the machine learning models trained on an email itself and its related data to determine if an email or a set of emails under analysis have potentially malicious characteristics. The cyber threat module can also factor this email characteristics analysis into its determination of the threat risk parameter.

[0070] The network module may have one or more machine learning models trained on the normal behavior of users, devices, and interactions between them, on a network that is tied to the email system. A user interface may have one or more windows to display network data and one or more windows to display emails and cyber security details about those emails through the same user interface on a display screen, which allows a cyber professional to pivot between network data and email cyber security details within one platform and consider them as an interconnected whole rather than separate realms on the same display screen.

[0071] The cyber-threat module can also factor this network analysis into its determination of the threat risk parameter.

[0072] The email module monitoring email activity and the network module monitoring network activity may both feed their data to a network & email coordinator module to correlate causal links between these activities to supply this input into the cyber-threat module. The cyber threat module can also factor this network activity link to a particular email causal link analysis into its determination of the threat risk parameter.

[0073] The cyber-threat defense system **100** uses various probes to collect activity, such as the user activity and the email activity, and then feeds that activity to the data store and, as needed to the cyber threat module and the machine learning models. The cyber threat module uses the collected data to draw an understanding of the email activity and user activity in the email system as well as update a training for the one or more machine learning models trained on this email system and its users. For example, email traffic can be collected by putting hooks into the e-mail application, such as Outlook or Gmail, and/or monitoring the internet gateway from which the e-mails are routed through. Additionally, probes may collect network data and metrics via one of the following methods: port spanning the organization's existing network equipment, inserting or re-using an in-line network tap, and/or accessing any existing repositories of network data (e.g., See FIG. **2**).

[0074] The cyber-threat defense system **100** may use multiple user interfaces. A first user interface may be constructed to present an inbox-style view of all of the emails coming in/out of the email system and any cyber security characteristics known of emails has a first window/column that displays the one or more emails under analysis and a second window/column with all of the relevant security characteristics known about that email or set of emails under analysis. The complex machine learning techniques determine anomaly scores that describe any deviation from normal that the email represents. These are rendered graphically in a familiar way that users and cyber professionals can recognize and understand.

[0075] The cyber-threat defense system **100** can then take actions to counter detected potential cyber threats. The autonomous response module, rather than a human taking action, can be configured to cause one or more rapid autonomous actions to be taken to contain the cyber threat when the threat risk parameter from the cyber threat module is equal to or above an actionable threshold. The cyber threat module's configured cooperation with the autonomous response module to cause one or more autonomous actions to be taken to contain the cyber threat improves computing devices in the email system by limiting the impact of the cyber threat from consuming unauthorized CPU cycles, memory space, and power consumption in the computing devices via responding to the cyber-threat without waiting for some human intervention. The cyber-threat defense system **100** may be hosted on a device, on one or more servers, and/or in its own cyber-threat appliance platform (e.g. see FIG. **2**).

[0076] FIG. **2** illustrates a block diagram of an embodiment of the cyber threat defense system monitoring email activity and network activity to feed this data to correlate causal links between these activities to supply this input into the cyber threat analysis. The network can include various computing devices such as desktop units, laptop units, smart phones, firewalls, network switches, routers, servers, databases, Internet gateways, the cyber-threat defense system **100**, etc.

[0077] The network module uses the probes to monitor network activity and can reference the machine learning models trained on a normal behavior of users, devices, and

interactions between them or the internet which is subsequently tied to the email system.

[0078] The user interface has both i) one or more windows to present/display network data, alerts, and events, and ii) one or more windows to display email data, alerts, events, and cyber security details about those emails through the same user interface on a display screen. These two sets of information shown on the same user interface on the display screen allows a cyber professional to pivot between network data and email cyber security details within one platform and consider them as an interconnected whole rather than separate realms.

[0079] The network module and its machine learning models are utilized to determine potentially unusual network activity in order to provide an additional input of information into the cyber threat module in order to determine the threat risk parameter (e.g. a score or probability) indicative of the level of threat.

[0080] A particular user's network activity can be tied to their email activity because the network module observes network activity, and the network & email coordinator module receives the network module observations to draw that into an understanding of this particular user's email activity to make an appraisal of potential email threats with a resulting threat risk parameter tailored for different users in the e-mail system. The network module tracks each user's network activity and sends that to the network & email coordinator component to interconnect the network activity and email activity to closely inform one-another's behavior and appraisal of potential email threats.

[0081] The cyber threat defense system **100** can now track possible malicious activity observed by the network module on an organization's network back to a specific email event observed by the e-mail module and use the autonomous response module to shut down any potentially harmful activity on the network itself, and also freeze any similar email activity triggering the harmful activity on the network.

[0082] The probes collect the user activity as well as the email activity. The collected activity is supplied to the data store and evaluated for unusual or suspicious behavioral activity, e.g. alerts, events, etc., which is evaluated by the cyber threat module to draw an understanding of the email activity and user activity in the email system. The collected data can also be used to potentially update the training for the one or more machine learning models trained on the normal pattern of life for this email system, its users and the network and its entities.

[0083] An example probe for the email system may be configured to work directly with an organization's email application, such as an Office 365 Exchange domain and receive a Blind Carbon Copy (BCC) of all ingoing and outgoing communications. The email module will inspect the emails to provide a comprehensive awareness of the pattern of life of an organization's email usage.

[0084] FIG. **3** illustrates a block diagram of an embodiment of the cyber threat module determining a threat risk parameter that factors in how the chain of unusual behaviors correlate to potential cyber threats and 'the likelihood that this chain of one or more unusual behaviors of the email activity and user activity under analysis fall outside of derived normal benign behavior;' and thus, is malicious behavior.

[0085] The user interface module **150** can graphically display logic, data, and other details that the cyber threat module goes through. The user interface module **150** displays an example email that when undergoing analysis exhibits characteristics, such as header, address, subject line, sender, recipient, domain, etc. that are not statistically consistent with the normal emails similar to this one. Thus, the user interface module **150** displays an example email's unusual activity that has it classified as a behavioral anomaly.

[0086] During the analysis, the email module can reference the one or more machine learning models that are self-learning models trained on a normal behavior of email activity and user activity associated with an email system. This can include various email policies and rules that are set for this email system. The cyber threat module may also reference the models that are trained on the normal characteristics of the email itself. The cyber threat module can apply these various trained machine learning models to data including metrics, alerts, events, meta data from the network module and the email module. In addition, a set of AI models may be responsible for learning the normal 'pattern of life' for internal and external address identities in connection with the rest of the network, for each email user. This enables the system to neutralize malicious emails which deviate from the normal 'pattern of life' for a given address identity for that user in relation to its past, its peer group, and the wider organization.

[0087] Next, the email module has at least a first email probe to inspect an email at the point it transits through the email application, such as Office 365, and extracts hundreds of data points from the raw email content and historical email behavior of the sender and the recipient. These metrics are combined with pattern of life data of the intended recipient, or sender, sourced from the data store. The combined set of the metrics are passed through machine learning algorithms to produce a single anomaly score of the email, and various combinations of metrics will attempt to generate notifications which will help define the 'type' of email.

[0088] Email threat alerts, including the type notifications, triggered by anomalies and/or unusual behavior of 'emails and any associated properties of those emails' are used by the cyber threat module to better identify any network events which may have resulted from an email borne attack.

[0089] In conjunction with the specific threat alerts and the anomaly score, the system may provoke actions upon the email designed to prevent delivery of the email or to neutralize potentially malicious content.

[0090] Next, the data store stores the metrics and previous threat alerts associated with each email for a period of time, for example this may be set to 27 days or more in one embodiment. This corpus of data is fully searchable from within the user interface module **150** and presents an invaluable insight into mail flow for email administrators and security professionals.

[0091] Next, the cyber threat module can issue an anomaly rating even when an unusual email does not closely relate to any identifiable malicious email. This value indicates how unusual the cyber threat module considers this email to be in comparison to the normal pattern of life for the organization and the specific internal user (either inbound recipient or outbound sender).

[0092] In one embodiment, the cyber threat module considers over 750 metrics and the organizational pattern of life for unusual behavior for a window of time. For example, the cyber threat module considers metrics and the organizational

pattern of life for unusual behavior and other supporting metrics for the past 7 days when computing the anomaly score, which is also factored into the final threat risk parameter.

[0093] FIG. **4** illustrates a block diagram of an embodiment of the cyber threat defense system referencing one or more machine learning models trained on gaining an understanding of a plurality of characteristics on an email itself and its related data including classifying the properties of the email and its metadata. The email module system extracts metrics from every email inbound and outbound. The user interface **150** can graphically display logic, data, and other details that the cyber-threat defense system goes through.

[0094] The cyber threat module in cooperation with the machine learning models analyzes these metrics in order to develop a rich pattern of life for the email activity in that email system. This allows the cyber threat module, in cooperation with the email module, to spot unusual anomalous emails that have bypassed/gotten past the existing email gateway defenses.

[0095] The email module detects emails whose content is not in keeping with the normal pattern of content as received by this particular recipient and passes the data to the cyber threat module for analysis. An example analysis may be as follows:

[0096] (i) To what level has the sender of this email been previously communicated with from individuals within the receiving organization?

[0097] (ii) How closely are the recipients of this mail related to those individuals who have previously communicated with the sender?

[0098] (iii) Is the content of this email consistent with other emails that the intended recipient sends or receives?

[0099] (iv) If any links or attachments present in the email were to be clicked or opened by the intended recipient, would this constitute anomalous activity for that individual's normal network behavior?

[0100] (v) Are the email properties consistent with this particular user's recent network activities?

[0101] Thus, the cyber threat module can also reference the machine learning models trained on an email itself and its related data to determine if an email or a set of emails under analysis have potentially malicious characteristics. The cyber threat module can also factor this email characteristics analysis into its determination of the threat risk parameter.

[0102] The email module can retrospectively process an email application's metadata, such as Office 365 metadata, to gain an intimate knowledge of each of their users, and their email addresses, correspondents, and routine operations. The power of the cyber threat module lies in leveraging this unique understanding of day-to-day user email behavior, of each of the email users, in relation to their past, to their peer group, and to the wider organization. Armed with the knowledge of what is 'normal' for a specific organization and specific individual, rather than what fits a predefined template of malicious communications, the cyber threat module can identify subtle, sophisticated email campaigns which mimic benign communications and locate threats concealed as everyday activity.

[0103] Next, the email module provides comprehensive email logs for every email observed. These logs can be filtered with complex logical queries and each email can be interrogated on a vast number of metrics in the email information stored in the data store.

[0104] Some example email characteristics that can be stored and analyzed are:

[0105] (i) Email direction: Message direction—outbound emails and inbound emails.

[0106] (ii) Send Time: The send time is the time and date the email was originally sent according to the message metadata.

[0107] (iii) Links: Every web link present in an email has its own properties. Links to web sites are extracted from the body of the email. Various attributes are extracted including, but not limited to, the position in the text, the domain, the frequency of appearance of the domain in other emails and how it relates to the anomaly score of those emails, how well that domain fits into the normal pattern of life of the intended recipient of the email, their deduced peer group, and their organization.

[0108] (iv) Recipient: The recipient of the email. If the email was addressed to multiple recipients, these can each be viewed as the 'Recipients.' The known identify properties of the email recipient, including how well known the recipient was to the sender, descriptors of the volume of mail, and how the email has changed over time, to what extend the recipient's email domain is interacted with inside the network.

[0109] (v) Subject: The email subject line.

[0110] (vi) Attachment: Every attachment associated with the message will appear in the user interface here as individual entries, with each entry interrogatable against both displayed and advanced metrics. These include, but are not limited to, the attachment file name, detected file types, descriptors of the likelihood of the recipient receiving such a file, descriptors of the distribution of files such of these in all email against the varying anomaly score of those emails.

[0111] (vii) Headers: Email headers are lines of metadata that accompany each message, providing key information such as sender, recipient, message content type for example.

[0112] The AI models may perform by the threat detection through a probabilistic change in normal behavior through the application of an unsupervised Bayesian mathematical model to detect behavioral change in computers and computer networks. The core threat detection system is termed the 'Bayesian probabilistic'. The Bayesian probabilistic approach can determine periodicity in multiple time series data and identify changes across single and multiple time series data for the purpose of anomalous behavior detection. From the email and network raw sources of data, a large number of metrics can be derived each producing time series data for the given metric.

[0113] The detectors in the cyber threat module including its network module and email module components can be discrete mathematical models that implement a specific mathematical method against different sets of variables with the target. Thus, each model is specifically targeted on the pattern of life of alerts and/or events coming from, for example, that cyber security analysis tool analyzing various aspects of the emails/coming from specific devices and/or users within a system, etc.

[0114] At its core, the cyber threat defense system **100** mathematically characterizes what constitutes 'normal' behavior in line with the normal pattern of life for that entity and organization based on the analysis of a large number/set of different measures of a device's network behavior. The

cyber threat defense system **100** can build a sophisticated 'pattern of life'—that understands what represents normality for every person, device, email activity, and network activity in the system being protected by the cyber threat defense system **100**.

[0115] The system may use a plurality of separate machine learning models. For example, a machine learning model may be trained on specific aspects of the normal pattern of life for entities in the system, such as devices, users, network traffic flow, outputs from one or more cyber security analysis tools analyzing the system, email characteristics etc.

[0116] Note, when the models leverage at least two different approaches to detecting anomalies: e.g. comparing each system's behavior to its own history, and comparing that system to its peers' history and/or e.g. comparing an email to both characteristics of emails and the activities and behavior of its email users, this multiple source comparison allows the models to avoid learning existing bad behavior as 'a normal' because compromised devices/users/components/emails will exhibit behavior different to their immediate peers.

[0117] In an embodiment, the one or more models may be trained on specific aspects of these broader concepts. For example, the models may be specifically trained on associations, attachments, compliances, data loss & transfers, general, meta data, hygiene, links, proximity, spoof, type, validation, and other anomalies.

[0118] Thus, for example, a first email model can retrospectively process Office 365 metadata to gain an intimate knowledge of users, email addresses, correspondents, and routine operations. Even in environments with encrypted email, the cyber defense system can derive key markers from metadata and provide valuable insights into correspondent identity, frequency of communication and potential risk.

[0119] The power of the cyber threat module lies in leveraging this unique understanding of day-to-day user email behavior in relation to their past, to their peer group, and to the wider organization. Armed with the knowledge of what is 'normal' for a specific organization and a specific email user, rather than what fits a predefined template of malicious communications, the cyber threat module can identify subtle, sophisticated email campaigns which mimic benign communications and locate threats concealed as everyday activity.

Defense System

[0120] FIG. **5** illustrates an example cyber threat defense system protecting an example network. The example network of FIG. **5** illustrates a network of computer systems **50** uses a threat detection system. The system depicted by FIG. **5** is a simplified illustration, which is provided for ease of explanation. The system **50** comprises a first computer system **10** within a building, which uses the threat detection system to detect and thereby attempts to prevent threats to computing devices within its bounds.

[0121] The first computer system **10** comprises three computers **1**, **2**, **3**, a local server **4**, and a multifunctional device **5** that provides printing, scanning and facsimile functionalities to each of the computers **1**, **2**, **3**. All of the devices within the first computer system **10** are communicatively coupled via a Local Area Network **6**. Consequently, all of the computers **1**, **2**, **3** are able to access the local server **4** via the LAN **6** and use the functionalities of the MFD **5** via the LAN **6**.

[0122] The LAN **6** of the first computer system **10** is connected to the Internet **20**, which in turn provides computers **1**, **2**, **3** with access to a multitude of other computing devices including server **30** and second computer system **40**. The second computer system **40** also includes two computers **41**, **42**, connected by a second LAN **43**.

[0123] In this exemplary embodiment of the present disclosure, computer **1** on the first computer system **10** comprises the threat detection system and therefore runs the threat detection method for detecting threats to the first computer system. As such, the computer system includes one or more processors arranged to run the steps of the process described herein, memory storage components required to store information related to the running of the process, as well as a network interface for collecting the required information. This method will now be described in detail with reference to FIG. **5**.

[0124] The computer **1** builds and maintains a dynamic, ever-changing model of the 'normal behavior' of each user and machine within the system **10**. The approach is based on Bayesian mathematics, and monitors all interactions, events, and communications within the system **10**—which computer is talking to which, files that have been created, networks that are being accessed.

[0125] For example, computer **2** is based in a company's San Francisco office and operated by a marketing employee who regularly accesses the marketing network, usually communicates with machines in the company's U.K. office in second computer system **40** between 9:30 AM and midday and is active from about 8:30 AM until 6 PM.

[0126] The same employee virtually never accesses the employee time sheets, very rarely connects to the company's Atlanta network, and has no dealings in South-East Asia. The threat detection system takes all the information that is available relating to this employee and establishes a 'pattern of life' for that person via the devices used by that person in the system, which is dynamically updated as more information is gathered. The 'normal' model of the pattern of life is used as a moving benchmark, allowing the system to spot behavior on a system that seems to fall outside of this normal pattern of life, and flags this behavior as anomalous, requiring further investigation.

[0127] The cyber defense self-learning platform uses machine-learning technology. The machine learning technology, using advanced mathematics, can detect previously unidentified threats, without rules, and automatically defend networks. Note, today's attacks can be of such severity and speed that a human response cannot happen quickly enough. Thanks to these self-learning advances, it is now possible for a machine to uncover emerging threats and deploy appropriate, real-time responses to fight back against the most serious cyber threats.

[0128] The threat detection system has the ability to self-learn and detect normality in order to spot true anomalies, allowing organizations of all sizes to understand the behavior of users and machines on their networks at both an individual and group level. Monitoring behaviors, rather than using predefined descriptive objects and/or signatures, means that more attacks can be spotted ahead of time and extremely subtle indicators of wrongdoing can be detected. Unlike traditional legacy defenses, a specific attack type or

new malware does not have to have been seen first before it can be detected. A behavioral defense approach mathematically models both machine and human activity behaviorally, at and after the point of compromise, in order to predict and catch today's increasingly sophisticated cyber-attack vectors. It is thus possible to computationally establish what is normal, in order to then detect what is abnormal.

[0129] This intelligent system is capable of making value judgments and carrying out higher value, more thoughtful tasks. Machine learning requires complex algorithms to be devised and an overarching framework to interpret the results produced. However, when applied correctly these approaches can facilitate machines to make logical, probability-based decisions and undertake thoughtful tasks.

[0130] Advanced machine learning is at the forefront of the fight against automated and human-driven cyber threats, overcoming the limitations of rules and signature-based approaches:

[0131] the machine learning learns what is normal within a network—it does not depend upon knowledge of previous attacks;

[0132] the machine learning thrives on the scale, complexity, and diversity of modern businesses, where every device and person is slightly different;

[0133] the machine learning turns the innovation of attackers against them—any unusual activity is visible;

[0134] the machine learning constantly revisits assumptions about behavior, using probabilistic mathematics; and

[0135] the machine learning is always up to date and not reliant on human input.

[0136] Utilizing machine learning in cyber security technology is difficult, but when correctly implemented it is extremely powerful. The machine learning means that previously unidentified threats can be detected, even when their manifestations fail to trigger any rule set or signature. Instead, machine learning allows the system to analyze large sets of data and learn a 'pattern of life' for what it sees.

[0137] Machine learning can approximate some human capabilities to machines, such as:

[0138] thought: it uses past information and insights to form its judgments;

[0139] real time: the system processes information as it goes; and

[0140] self-improving: the model's machine learning understanding is constantly being challenged and adapted, based on new information.

[0141] New unsupervised machine learning therefore allows computers to recognize evolving threats, without prior warning or supervision.

Unsupervised Machine Learning

[0142] Unsupervised machine learning works things out without pre-defined labels. In the example of sorting a series of different entities, such as animals, the system would analyze the information and work out the different classes of animals. This allows the system to handle the unexpected and embrace uncertainty when new entities and classes are examined. The system does not always know what it is looking for but can independently classify data and detect compelling patterns.

[0143] The cyber threat defense system's unsupervised machine learning methods do not require training data with pre-defined labels. Instead, they are able to identify key patterns and trends in the data, without the need for human input. The advantage of unsupervised learning in this system is that it allows computers to go beyond what their programmers already know and discover previously unknown relationships.

[0144] The cyber threat defense system uses unique implementations of unsupervised machine learning algorithms to analyze network data at scale, intelligently handle the unexpected, and embrace uncertainty. Instead of relying on knowledge of past threats to be able to know what to look for, it is able to independently classify data and detect compelling patterns that define what may be considered to be normal behavior. Any new behaviors that deviate from those, which constitute this notion of 'normality,' may indicate threat or compromise. The impact of the cyber-threat defense system's unsupervised machine learning on cyber security is transformative:

[0145] threats from within, which would otherwise go undetected, can be spotted, highlighted, contextually prioritized and isolated using these algorithms;

[0146] the application of machine learning has the potential to provide total network visibility and far greater detection levels, ensuring that networks have an internal defense mechanism; and

[0147] machine learning has the capability to learn when to action automatic responses against the most serious cyber threats, disrupting in progress attacks before they become a crisis for the organization.

[0148] This new mathematics not only identifies meaningful relationships within data, but also quantifies the uncertainty associated with such inference. By knowing and understanding this uncertainty, it becomes possible to bring together many results within a consistent framework—the basis of Bayesian probabilistic analysis. The mathematics behind machine learning is extremely complex and difficult to get right. Robust, dependable algorithms are developed, with a scalability that enables their successful application to real-world environments.

Overview

[0149] In an embodiment, a closer look at the cyber threat defense system's machine learning algorithms and approaches is as follows.

[0150] The cyber threat defense system's probabilistic approach to cyber security is based on a Bayesian framework. This allows it to integrate a huge number of weak indicators of potentially anomalous network behavior to produce a single clear measure of how likely a network device is to be compromised. This probabilistic mathematical approach provides an ability to understand important information, amid the noise of the network—even when it does not know what it is looking for.

Ranking Threats

[0151] Advantageously, the cyber threat defense system's approach accounts for the inevitable ambiguities that exist in data and distinguishes between the subtly differing levels of evidence that different pieces of data may contain. Instead of generating the simple binary outputs 'malicious' or 'benign,' the cyber threat defense system's mathematical algorithms produce outputs that indicate differing degrees of potential compromise. This output enables users of the system to rank different alerts in a rigorous manner and prioritize those that

most urgently require action, simultaneously removing the problem of numerous false positives associated with a rule-based approach.

[0152] The cyber threat defense system mathematically characterizes what constitutes 'normal' behavior based on the analysis of a large number/set of different measures of a devices network behavior, examples include:

[0153] server access;

[0154] data access;

[0155] timings of events;

[0156] credential use;

[0157] DNS requests; and

[0158] other similar parameters.

[0159] Each measure of network behavior is then monitored in real time to detect anomalous behaviors.

Clustering

[0160] To be able to properly model what should be considered as normal for a device, its behavior must be analyzed in the context of other similar devices on the network. To accomplish this, the cyber threat defense system leverages the power of unsupervised learning to algorithmically identify naturally occurring groupings of devices, a task which is impossible to do manually on even modestly sized networks.

[0161] In order to achieve as holistic a view of the relationships within the network as possible, the cyber threat defense system simultaneously employs a number of different clustering methods including matrix based clustering, density based clustering and hierarchical clustering techniques. The resulting clusters are then used to inform the modeling of the normative behaviors of individual devices.

[0162] Clustering: At a glance:

[0163] analyzes behavior in the context of other similar devices on the network;

[0164] algorithms identify naturally occurring groupings of devices—impossible to do manually; and

[0165] simultaneously runs a number of different clustering methods to inform the models.

Network Topology

[0166] Any cyber threat detection system preferably recognizes that a network is far more than the sum of its individual parts, with much of its meaning contained in the relationships among its different entities, and that complex threats can often induce subtle changes in this network structure. To capture such threats, the cyber-threat defense system employs several different mathematical methods in order to be able to model multiple facets of a networks topology.

[0167] One approach is based on iterative matrix methods that reveal important connectivity structures within the network. In tandem with these, the cyber threat defense system has developed innovative applications of models from the field of statistical physics, which allow the modeling of a network's 'energy landscape' to reveal anomalous substructures that may be concealed within.

Network Structure

[0168] A further important challenge in modeling the behaviors of network devices, as well as of networks themselves, is the high-dimensional structure of the problem with the existence of a huge number of potential predictor variables. Observing packet traffic and host activity within an enterprise LAN, WAN and Cloud is difficult because both input and output can contain many inter-related features (protocols, source and destination machines, log changes and rule triggers, etc.). Learning a sparse and consistent structured predictive function is crucial to avoid the curse of over fitting.

[0169] In this context, the cyber threat defense system has employed a cutting edge large-scale computational approach to learn sparse structure in models of network behavior and connectivity based on applying L1-regularization techniques (e.g. a lasso method). This allows for the discovery of true associations between different network components and events that can be cast as efficiently solvable convex optimization problems and yield parsimonious models.

Recursive Bayesian Estimation

[0170] The unsupervised machine learning methods can use a probabilistic approach based on a Bayesian framework. The machine learning allows the CSA **100** to integrate a huge number of weak indicators/low threat values by themselves of potentially anomalous network behavior to produce a single clear overall measure of these correlated anomalies to determine how likely a network device is to be compromised. This probabilistic mathematical approach provides an ability to understand important information, amid the noise of the network—even when it does not know what it is looking for.

[0171] The CSA **100** can use a Recursive Bayesian Estimation. To combine these multiple analyses of different measures of network behavior to generate a single overall/comprehensive picture of the state of each device, the threat defense system takes advantage of the power of Recursive Bayesian Estimation (RBE) via an implementation of the Bayes filter.

[0172] Using RBE, the cyber threat defense system's mathematical models are able to constantly adapt themselves, in a computationally efficient manner, as new information becomes available to the system. They continually recalculate threat levels in the light of new evidence, identifying changing attack behaviors where conventional signature-based methods fall down.

[0173] Training a model can be accomplished by having the model learn good values for all of the weights and the bias for labelled examples created by the system. In some cases, the system may start with no labels initially. A goal of the training of the model can be to find a set of weights and biases that have low loss, on average, across all examples.

[0174] An anomaly detection technique that can be used is supervised anomaly detection that requires a data set that has been labelled as "normal" and "abnormal" and involves training a classifier. Another anomaly detection technique that can be used is an unsupervised anomaly detection that detects anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal, by looking for instances that seem to fit least to the remainder of the data set. The model representing normal behavior from a given normal training data set can detect anomalies by establishing the normal pattern and then test the likelihood of a test instance under analysis to be generated by the model. Anomaly detection can identify rare items, events or observations which raise suspicions by

differing significantly from the majority of the data, which includes rare objects as well as things like unexpected bursts in activity.

Likely Recipient Classifier

[0175] In a further aspect of the present disclosure, a cyber threat risk detection platform of the cyber threat protection system may be configured to analyze draft emails prior to sending by a user of the system. In particular, an apparatus **160** may comprise a Likely Recipient Classifier **162**, one or more machine learning models **164**, an autonomous response module **166**, and a user interface **168**.

[0176] The apparatus **160** may be configured to process the draft email/email for sending in order to verify or otherwise authenticate a recipient of one or more recipients entered into the "Recipient" field/"To" field of the draft email. In this manner, the apparatus **160** may act to reduce the likelihood of an outbound email being sent to an unintended recipient and to prevent any associated cyber threat risk, for example the unintended disclosure of confidential or otherwise sensitive information.

[0177] As discussed above, this aspect of the present disclosure may be applied to any structured document that is to be sent by a sender to an addressed recipient. While the following disclosure will be discussed in relation to the example of emails and an email system, further examples of structured documents that the present disclosure also applies to include, but are not limited to, instant messages, text messages and other structured electronic communications.

[0178] Accordingly, in one embodiment a Likely Recipient Classifier **162** may be arranged to analyze a plurality of emails that have been sent from a sending user to a given recipient. This given recipient may be an individual email address, or alternatively the recipient may be considered to be any email address within a given domain—for example, any email address associated with a given company/organization. The given recipient may be considered to be a label/class/category identifying the outbound emails that are sent to that recipient email address. A plurality of metrics/characteristics may be extracted from each email (including its related data). The one or more machine learning models **164** may then be trained to identify which particular metrics/characteristics of an email (including any related data) can be considered, when taken alone or in combination, to be strong indicators that the email has been or will be addressed to a given recipient.

[0179] The set of strong indicators may be characteristics that are common in emails addressed to the given recipient but rarely found in emails addressed to alternative recipients, or alternatively rare characteristics that are not found for alternative recipients, and they may be referred to as "key indicators". The more often/higher the frequency of occurrence of these characteristics in emails addressed to the given recipient, then the more weight is given to those specific characteristics in the key indicators. Similarly, the rarer the characteristics are, then the more weight is given to those specific characteristics in the key indicators.

[0180] Once the one or more machine learning models **164** have been trained on the past emails sent by the sender to various recipients, these trained models may be used by the Likely Recipient Classifier **162** to parse and process characteristics extracted from a previously unseen draft email to be sent by the sender in order to infer an expected recipient of that draft email. This processing may check for matches (for example using k-means clustering) between the characteristics extracted from the draft email and the key indicators that have been identified for the recipients that the sender has previously sent emails to and are therefore known recipients to the sender. By considering each of these known recipients in turn and determining a score associated with the probability of a match between the characteristics extracted from the draft email and the key identifiers for each known recipient, the Likely Recipient Classifier may determine the known recipient with the highest probability of a match for the draft email, who may be referred to as the expected recipient.

[0181] The autonomous response module **166** may obtain the expected recipient (having the highest determined probability of a match) from the Likely Recipient Classifier **162** and, where the sender has not yet indicated a recipient in the "To" field of the email, the autonomous response module may cause an alert to be displayed to the sender on the user interface **168**, where the alert provides a recommendation or suggestion for the expected recipient to be entered into the "To" field of the draft email.

[0182] Where at least one recipient has already been indicated by the sender in the "To" field of the draft email, the autonomous response module **166** may compare the probability of a match for the draft email with the indicated recipient and the expected recipient. If the indicated recipient is determined to have a probability score that is lower than the expected recipient, then the autonomous response module may cause an alert to be displayed to the sender on the user interface **168**, where the alert indicates that the expected recipient has been determined to have a better match to the email characteristics than the indicated recipient. This alert may also provide a recommendation or suggestion for the indicated recipient to be replaced with the expected recipient in the "To" field of the draft email. The sender can then review this alter and decide whether to accept the recommendation or whether to ignore/dismiss the alert. If the alert recommendation is accepted, then the autonomous response module may cause the expected recipient to replace the indicated recipient in the "To" field of the draft email without further input from the sender/user. If the alert recommendation is dismissed, then this dismissal may be recorded in a log associated with the autonomous response module **166**.

[0183] The alert recommendation may also be accompanied by an indication of the logic why the Likely Recipient Classifier **162** has provided the recommendation, for example one or more of the relevant key identifiers may be displayed to the sender user.

[0184] In some embodiments, where an alert recommendation is generated, the apparatus **160** may be configured to prevent the draft email from being sent until the sender has either accepted or dismissed the alert recommendation.

[0185] The apparatus **160** may be configured to process the draft email while continually or periodically while it is being written. Alternatively, the processing of the draft email may be deferred until the draft email has been completed and it is ready to send.

[0186] The metrics extracted from the emails by the apparatus **160** may include characteristics relating to aspects of the language content of the email, such as the constituent words and/or phrases in a body text of the email, or in the subject line, the name/file type/or content of an attachment if present.

[0187] By performing this machine learning model training for emails from a given sender to a plurality of recipients, the one or more machine learning models can learn and refine/filter these key indicators to enable the one or more machine learning models **164** to identify the words or terms that the particular sender frequently uses in emails to one recipient relative to the other recipients that the sender has previously sent emails to. In this manner, the training may filter out words or terms that the sender uses frequently with many or with a majority of recipients, since these words would not be strong or key indicators for any one particular recipient.

[0188] Once the characteristics have been extracted from the email and its related data, a first pre-processing step may be performed to reduce the words to their word stem. Stemming inflected words in this manner results in an improved efficiency of training and subsequent classification because the total number of word stems from an email will typically be less than the total number of different words present in the unprocessed email. However, this stemming process has also been found to improve the accuracy of the classification since this prevents the use of a variety of inflections of a given word stem from reducing the relative strength of that word stem/set of word inflections in being an effective key indicator for the identity of the intended recipient of an email written by the given sender.

[0189] The training data set preferably includes all emails that the sender has sent to any recipient within a given timeframe, for example this may include all emails sent by the sender that are still accessible to the apparatus **160**, taking into account document retention policies. Advantageously, the training of the one or more machine learning modules may also be updated periodically with further training data, i.e. including the emails sent by the sender since the last update. This enlarged data set serves to improve the classification accuracy of the Likely Recipient Classifier **162** and also enables the apparatus **160** to adapt to any changes in the email style of the sender. In one example embodiment, this updated training may be performed daily; however it will be appreciated that alternative periodicities may also be selected without departing from the scope of the present disclosure.

[0190] While training, and updating, the Likely Recipient Classifier **162** on a set of training data corresponding to emails sent by a single sender results in an accurate classifier for further emails sent by that sender, this also serves to limit the potential size of the training data set. This may result in some words being identified as being a key indicator for a given recipient (based on the relatively small sample size of the past emails sent by the sender) when in reality the word is actually too common to be a key indicator that can reliably identify any particular recipient.

[0191] In order to combat this, the training of the one or more machine learning models may include filtering out words or terms that have been identified as common from a much larger (generic) training data set of emails that is not limited to those sent by any individual sender. For example, this larger generic training data set may include hundreds of thousands of previous emails. For performing this filtering processing of the words/key indicators during the periodic training of the machine learning models, an extensive bloom filter (or set of bloom filters) may be used to enable the apparatus **160** to make an efficient probabilistic determination of whether a given word or key identifier identified during the training is in the set(s) of common words determined from the generic training data set.

[0192] However, the inventors have also appreciated that it might be desirable to not filter out some of the terms that appear fairly common from the generic training data set. For example, while some terms might be used in emails addressed to a number of different recipients, in some cases these recipients may all be in the same industry and accordingly the term may actually be a strong indicator for that industry. When included in the key indicators for a given sender, such industry specific terms may still contribute to the accuracy of the classification. In order to achieve this, the relatedness of different organizations/domains within the generic training data set may be taken into account when determining whether such terms should be filtered out to prevent them from being used as key indicators.

[0193] In some embodiments, the alert recommendation may only be displayed to the sender if the probability score for the indicated recipient is at least a certain amount lower than the probability score for the expected recipient, i.e. if the probability score for the expected recipient is at least a certain amount higher than the probability score for the indicated recipient. This threshold amount may be set to prevent false alerts from being displayed to the sender in the event that there is only a marginal difference in the relative probability scores for the match.

[0194] Further Likely Recipient Classifiers may be trained for other senders, some of whom may be associated with the same network or organization as the given sender/user. In some embodiments, the Likely Recipient Classifier **162** may take into account key indicator information from other senders that are from the same network as the user when determining the alert to be displayed to the user for a given draft email. This may improve the alert recommendations provided to the user by considering terms or words that the user in question may not have used (or at least not commonly) for a particular recipient in the past, but that others on the network have sent to a particular indicated recipient. This may be used to avoid wrongly suggesting the user's indicated recipient should be changed. This may also be used to propose a recipient that the particular user has never messaged, but others on their network have, especially if the machine learning model training indicates that the term(s) or word(s) are unique to that unknown recipient.

[0195] In a further embodiment, the characteristics extracted from the emails for training the one or more machine learning models **164** and for classifying unseen emails by the Likely Recipient Classifier **162** may include the other recipients of the email/other indicated recipients of the draft email, respectively. In this manner, common groups of recipients that are frequently seen together in emails sent by the sender may be taken into account when determining the corresponding alert recommendation.

[0196] A computer implemented method for determining and acting on a cyber threat risk of an email to be sent from a sender to an indicated recipient may also be provided in accordance with the present disclosure. At step **S1**, the email may be received and parsed at the Likely Recipient Classifier **162** to extract a plurality of characteristics from it. At step **S2**, the Likely Recipient Classifier **162** may access and use one or more machine learning models **164** that have been trained on the classification of emails with one or more of a plurality of recipients known to the sender, based on a plurality of characteristics of the emails. In particular, the

Likely Recipient Classifier **162** classifies the email by determining a set of respective match probability scores between the extracted plurality of characteristics and each of the known recipients, including the indicated recipient. At step S**3**, the autonomous response module **166** determines an expected recipient corresponding to the known recipient having the highest probability of a match. Then at step S**4**, if the indicated recipient is not the expected recipient, the autonomous response module **166** causes an alert to be displayed to the sender on the sender user interface **178** indicating that the expected recipient has a higher probability of a match with the email than the indicated recipient.

[0197] Further embodiments of this computer implemented method are set out in clauses 23 to 33 below.

Inducement Classifier

[0198] FIG. **7**A illustrates an apparatus according to another aspect of the present disclosure. The cyber threat detection system may be configured to analyze incoming emails received at the cyber threat detection system in order to detect and act upon emails that are classified as being malign or malicious. In particular, an apparatus **170** may comprise an Inducement Classifier **172**, one or more machine learning models **174**, an autonomous response module **176**, and a user interface **178**.

[0199] The apparatus **170** may be configured to parse and process incoming emails for detecting when an email attempts, by its content (including any embedded links, attachments, photos etc.), to 'induce' a certain behavior in the recipient of the email. In this manner, the apparatus **170** may act to identify malign emails such that one or more automated actions may be performed to neutralize any associated cyber threat risk.

[0200] As discussed above, this aspect of the present disclosure may be applied to any structured document with defined fields that has been sent to the user by a sender. While the following disclosure will be discussed in relation to the example of emails and an email system, further examples of structured documents that the present disclosure also applies to include, but are not limited to, instant messages, text messages and other structured electronic communications.

[0201] Accordingly, in one embodiment an Inducement Classifier **172** may be arranged to analyze an email that has been sent to the user by a given email sender to extract a plurality of characteristics from the email, and to classify the email based upon the extracted plurality of characteristics. The fields of the email that are analyzed may include the body text of the email, the subject line, the sender field, attachment names, etc. The plurality of characteristics preferably relates to the tone and/or content of the email such that the email can be classified into one of a plurality of categories of email type. The categories of email type may relate to the type of behavior requested of the recipient by the email.

[0202] In one embodiment, the Inducement Classifier **172** achieves the analysis and subsequent classification of the email under analysis using a multiple-pronged approach comprising:

[0203] considering extracted characteristics relating to the content of the email using word analysis;

[0204] considering extracted characteristics relating to the tone of the email using structure analysis; and

[0205] considering the type of induced behavior that is determined to be requested of the recipient.

[0206] However, it will be appreciated that similar classification could be achieved using only word analysis for example. The word analysis aims to identify key words or phrases that may be associated with each inducement category, but that rarely occur in emails that would not be categorized as malign. For example this may be by comparing the relative frequency density of such words and phrases in respective emails. The words and phrases considered may include the name and/or content of an attachment to the email if one exists.

[0207] In one example, the frequency density may be compared for one word phrases, two word phrases, three word phrases, or any combination of these. Each of the words in these phrases are preferably pre-processed to reduce the words to their corresponding word stem. Stemming inflected words in this manner results in an improved efficiency of training and subsequent classification in the apparatus **170** because the total number of word stems from an email will typically be less than the total number of different words present in the unprocessed email. However, this stemming process has also been found to improve the accuracy of the classification since this stemming focusses the analysis on the core meaning of the word phrase rather than any distribution of the various inflections of a given word stem used.

[0208] In one embodiment, this may be achieved using a k-means model applied to a modified Term Frequency—Inverse Document Frequency extraction from certain subsets of this metadata to determine how important a specific word, feature or phrase is.

[0209] The one or more machine learning models **174** may be trained using a training data set comprising example emails that have been labelled with each of the categories to be classified.

[0210] The Inducement Classifier **172** may be an ensemble of instructions that takes in the characteristics/metrics and then outputs a prediction about the category of the email being analyzed. The Inducement Classifier **172** may use a hypothesis learned by a machine learning algorithm or a discrete-valued function.

[0211] In one embodiment, the malign categories of emails that may represent a cyber threat risk may include one or more of "extortion," "phishing," "solicitation," "commercial spam," "other spam" and "other."

[0212] For example, an extortion email typically uses words and/or links in the body of the email to attempt to induce fear and/or embarrassment in the recipient and includes words directed to the extraction, typically by blackmail, of some sort of payment from the email recipient.

[0213] In another example, a solicitation email typically uses words in the email body text to cause or encourage the recipient to purchase, for the benefit of the sender of the email, a product or service discussed in the body of the email or in an embedded link in the email.

[0214] Once the one or more machine learning models **174** have been trained on the labelled example email training data set(s), these trained models may be used by the Inducement Classifier **172** to parse and process characteristics extracted from a previously unseen incoming email received by the user in order to infer the category that should be associated with that incoming email, along with a score

associated with the probability of a match between the email characteristics and the respective categories.

[0215] The autonomous response module **176** may then obtain the probability score associated with the strength of the match of the incoming email to each of the possible malign/inducement categories. If the associated probability score determined for the one or more malign categories (or an average across multiple categories) is above a threshold, the autonomous response module **176** may be configured to cause one or more autonomous actions to be taken in order to contain or neutralize the malign nature of the incoming email. For example, in one embodiment one or more actions may be taken if the probability score is 60% or higher. Further, more aggressive, actions may be configured to be taken if the probability score exceeds higher thresholds.

[0216] The autonomous response module **176** may have a library of possible response action types and specific actions that the autonomous response module is capable of. These may include focused response actions selectable through the user interface **178** that are contextualized to autonomously act on specific email elements of a malicious email, rather than a blanket quarantine or block approach on that email, to avoid business disruption to a particular user of the email system.

[0217] The autonomous response module is able to take measured, varied actions towards those email communications to minimize business disruption in a reactive, contextualized manner. In this manner, the autonomous response module works to neutralize malicious emails, and deliver preemptive protection against targeted, email-borne attack campaigns in real time.

[0218] Optionally, an indication of the logic regarding why the Inducement Classifier **172** has provided categorized the email in the way that is has may be displayed to the user on the user interface **178**, for example one or more of the relevant words or terms may be displayed to the user.

[0219] FIG. **7**B illustrates a diagram of an embodiment of a segmentation module employed by a phishing site detector module to analyze key text-like features of an example legitimate site and multiple example legitimate sites. The segmentation module breaks up an image of a page of a site under analysis into multiple segments and then analyze each segment of the image of that page to determine visually whether a key text-like feature exists in that segment. A signature creator creates a digital signature for each segment containing a particular key text-like feature. The digital signature for that segment containing the particular key text-like feature at least is indicative of a visual appearance of the particular key text-like feature. One or more trained AI models trained to compare digital signatures from a set of key text-like features detected in the image of that page of the unknown site under analysis to digital signatures of a set of key text-like features from a plurality of known bad phishing sites in order to output at least a likelihood of maliciousness of the unknown site under analysis. Note, any portions of the cyber security appliance **100** implemented as software can be stored in one or more non-transitory memory storage devices in an executable format to be executed by one or more processors.

[0220] The phishing site detector in the cyber security appliance **100** uses machine learning approaches 1) to detect segments of text-like features in a screenshot of a page under analysis in order 2) to create a digital signature for the text-like features, 3) to transform each segment of the screen shot into a comparable rendered size of the similar feature represented in the library of digital signatures, 4) to classify these text-like features in the segments into likely subject matter categories (keywords like 'email' or brand names such as 'Microsoft'), and 5) to compare the existing library of feature digital signatures with those derived for newly created segments of a page with a corresponding text-like feature from a site the phishing site detector puts under analysis.

[0221] The segmentation module may break up an image of a page of a site under analysis into multiple segments of the image of the page of the unknown site under analysis. The page of the unknown site under analysis can be a log-in page for the unknown site. The segmentation module can use a machine learning algorithm for breaking up and segmenting the image of the site under analysis. The segmentation module can analyze each segment of the image of that page to determine visually whether a key text-like feature exists in that segment. The segmentation module both detects a set of key text-like features in the multiple segments of the image and determines coordinates around each key text-like feature in that set of key text-like features.

[0222] The segmentation module applying the machine learning algorithm identifies areas of key features along with their coordinates on the image of the page, (e.g., in the visual appearance of the site) as rendered on the end user's computing device. The segmentation module forms a bounding box around each of these key features. The phishing site detector uses the literal visual representation of the key feature, such as a word/phrase, to detect it, so it is not 'trickable' in the same way that pure OCR is. Words, phrases, logos, etc. come in different fonts, colors, and styles of text as well as misspelled word can be used in a phishing site that visually with a quick glance appear to an actual word that the user would be expecting to see, such as 'sign 1n' with a one. OCR merely ascertains letters themselves and nothing about the visual appearance of the letters.

[0223] FIG. **8** illustrates a block diagram of an embodiment of example autonomous actions that the autonomous response module can be configured to take without a human initiating that action. If the processing determines that a threshold has not been exceeded, then the incoming email may proceed without further action by the autonomous response module **176**. However, if a threshold is exceeded then one or more of the appropriate action types may be initiated. Example actions may include one or more of the following:

[0224] Hold Message: The autonomous response module may hold the email an prevent delivery to the user or their email inbox, for example due to suspicious content or attachments. Held emails can be reprocessed and released by a user after investigation. If delivery has already been performed, then the email may be removed from the user's inbox. The original mail may be maintained in a buffered cache by the data store and can be recovered, or sent to an alternative mailbox, using a 'release' button in the user interface **178**.

[0225] Lock Links: The autonomous response module replaces the URL of a link such that a click of that link will first divert the user via an alternative destination. The alternative destination may optionally request confirmation from the user before proceeding. The original link destination and original source may be subject to additional checks before the user is permitted to access the source.

[0226] Convert Attachments: The autonomous response module converts one or more attachments of the email to a safe format, for example flattening the attachment file by converting it into a PDF using image conversion. This delivers the content of the attachment to the intended recipient, but with vastly reduced risk since aspects such as macros or other automated scripts may be removed by the conversion. For attachments which are visual in nature, such as images, pdfs and Microsoft Office formats, the attachments may be processed into an image format and subsequently rendered into a PDF (in the case of Microsoft Office formats and PDFs) or into an image of the original file format (if an image). In some email systems, the email attachment may be initially removed and replaced with a notification informing the user that the attachment is undergoing processing. When processing is complete the converted attachment may be inserted back into the email.

[0227] Double Lock Links: The autonomous response module replaces the URL with a redirected email link. If the link is clicked, the user will be presented with a notification to that user that they are not permitted to access the original destination of the link. The user will be unable to follow the link to the original source, but their intent to follow the link may be recorded by the data store via the autonomous response module for a subsequent follow up with the user.

[0228] Strip Attachments: The autonomous response module strips one or more attachments of this email. Most file formats are delivered as converted attachments; file formats which do not convert to visible documents (e.g. executables, compressed types) are stripped to reduce risk. The 'Strip attachment' action will cause the system to remove the attachment from the email and replace it with a file informing the user that the original attachment was removed.

[0229] Junk action: The autonomous response module classifies the email as junk or other malicious email and diverts it to the user's junk folder, or other nominated destination such as 'quarantine.'

[0230] Redirect: The autonomous response module may ensure the email is not delivered to the user but is instead diverted to a specified email address.

[0231] Copy: The autonomous response module may ensure the email is delivered to the original recipient, but that a copy is also sent to another specified email address.

[0232] Do not hold or alter: For particular users, the autonomous response module may be configured to ignore actions that would otherwise be taken for other users.

[0233] Take no action on attachments: For particular users, the autonomous response module may override any attachment actions that would be otherwise taken.

[0234] Header and body action: The autonomous response module may insert specific/custom text into the email body or subject line to add to or replace existing text, images, or other content in a header and/or body of the email.

[0235] Unspoof: The autonomous response module may identify standard email header address fields (e.g. rfc822 type) and replace the personal name and the header email address with an alternative name or email address which might reveal more about the true sender of the email. This mechanism significantly reduces the psychological impact of spoof attempts.

[0236] In one embodiment, the apparatus **170** may be configured to further take into account a variety of structural metrics/characteristics for use in the training and subsequent inference stages. These structural characteristics may consider factors such as one or more of: the average length of sentences and/or paragraphs used in the email body text (derived ratio analysis), the density of numbers or capitalization in the text, the presence of phone numbers, email addresses, large round numbers (may be linked to extortion for example), words that are a combination of character types (e.g. mixing Latin letters with either numbers or non-Latin characters, such as Cyrillic), currency values (in particular cryptocurrencies) and the format of the addressing fields.

[0237] For example, an extortion/blackmail email is typically long in content with details of the blackmail compared to a solicitation email, which subsequently highlights the products or service being offered. On the contrary, a legitimate work-related email rarely requests money from the recipient unless contract negotiations are being discussed. However, the general tone and induced behavior of a work-related contract negotiation is quite different than that of an extortion message. Moreover, attached contract terms for such a work-related email may be expected to be very lengthy compared to an extortion email.

[0238] In the above, the Inducement Classifier **172** and the one or more machine learning models **174** may extract and analyze may different metrics/characteristics for each incoming email. In one specific example over 180 characteristics may be used for the word analysis and over 120 characteristics may be used for the structure analysis.

[0239] An overarching classification model may also be fitted using these characteristics and used to score an incoming email in the various inducement categories, which in turn may be used to calculate an overall inducement "badness" score associated with the email. In an embodiment, this overarching classifier may use a different set of training data to that used for the word analysis.

[0240] Where a given user has a history of emails from a legitimate contact that are not classified as inducement, but the user suddenly starts to receive one or more emails from the contact that are classified as inducement by the Inducement Classifier **172**, then it may be inferred that the contact's email account has been hijacked or otherwise compromised. In such a situation, the system may optionally send a notification to the contact to alert them to this possible hijacking/compromise so that appropriate action may be taken.

[0241] In a further embodiment a secondary classifier may be trained on the analysis of emails based only on the structural characteristics (i.e. not also on the word characteristics). This may be particularly beneficial in circumstances where it is determined that an incoming email is written in a language that the one or more machine learning models have not been trained on, since the relevance of the structural characteristics will still hold true and may enable the system and method to calculate inducement scores of emails that are written in these languages unknown to the Inducement Classifier **172**. The language may be detected, for example, using a language classifier trained to identify the language of a given piece of text.

[0242] This highlights the additional benefit of considering the structure of an email as well as its word content when identifying emails that may be designed to induce certain bad behaviors in a receiving user, particularly because it has been appreciated that the structure of a typical solicitation, phishing, extortion etc. email may be preserved even when the language is translated. Furthermore, the emphasis on

structure may also allow for topical spear-phishing (such as topical emails relating to Covid-19) in both English and Non-English emails to be identified since the classifier would not rely solely on the meaning of topical terms, which may not have been present in the training datasets.

[0243] In some embodiments, the user interface **178** may be provided with an administrative tool for the user to set which types of autonomous actions the autonomous response module is configured to perform and for setting the relevant thresholds for triggering these autonomous actions. The types of actions and specific actions that the autonomous response module **176** may be customized with may be set individually for different users and/or different parts of the system.

[0244] In one embodiment, the autonomous response module may initially be run in human confirmation mode in which all autonomous, intelligent interventions are confirmed by a human operator. As the apparatus **170** refines and nuances its understanding of an organization's email behavior, the level of autonomous action can be increased until eventually no human supervision is required for each autonomous response action. Most security teams will spend very little time in the user interface administrative tool once this level is reached. At that stage, the autonomous response module action may respond to and neutralize malicious emails without the need for any active management.

[0245] The intelligent system of the present disclosure is capable of making value judgments and carrying out higher value, more thoughtful tasks. Machine learning requires complex algorithms to be devised and an overarching framework to interpret the results produced. However, when applied correctly these approaches can facilitate machines to make logical, probability-based decisions and undertake thoughtful tasks.

[0246] This new mathematics not only identifies meaningful relationships within data, but also quantifies the uncertainty associated with such inference. By knowing and understanding this uncertainty, it becomes possible to bring together many results within a consistent framework—the basis of Bayesian probabilistic analysis. The mathematics behind machine learning is extremely complex and difficult to get right. Robust, dependable algorithms are developed, with a scalability that enables their successful application to real-world environments.

[0247] A computer implemented method for determining and acting on a cyber threat risk of an email that has been sent to a user from a given sender may also be provided in accordance with the present disclosure. At step S**10**, the email may be received and parsed at the Inducement Classifier **172** to extract a plurality of characteristics from it. At step S**11**, the Inducement Classifier **172** may access and use one or more machine learning models **174** that have been trained on the classification of emails with one or more of the plurality of malign categories based on the extracted characteristics of emails, in particular, the Inducement Classifier **172** classifies the received email with one or more of the plurality of malign categories based on the extracted plurality of characteristics and the one or more machine learning models **174**. At step S**12**, a score is determined for the probability of a match between the email characteristics and each of the plurality of malign categories. Then at step S**13**, the autonomous response module **176** determines whether one of these scores is above a threshold; if the score determined for the one or more malign categories is above the threshold, the autonomous response module **176** causes one or more actions to contain the malign nature of the received email to be initiated as discussed above.

[0248] Further embodiments of this computer implemented method are set out in clauses 35 to 40 below.

[0249] The method, apparatus and system are arranged to be performed by one or more processing components with any portions of software stored in an executable format on a computer readable medium. Thus, any portions of the method, apparatus and system implemented as software can be stored in one or more non-transitory memory storage devices in an executable format to be executed by one or more processors. The computer readable medium may be non-transitory and does not include radio or other carrier waves. The computer readable medium could be, for example, a physical computer readable medium such as semiconductor memory or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disc, and an optical disk, such as a CD-ROM, CD-R/W or DVD.

[0250] The various methods described above may be implemented by a computer program product. The computer program product may include computer code arranged to instruct a computer to perform the functions of one or more of the various methods described above. The computer program and/or the code for performing such methods may be provided to an apparatus, such as a computer, on a computer readable medium or computer program product. For the computer program product, a transitory computer readable medium may include radio or other carrier waves.

[0251] A cloud provider platform may include one or more of the server computing systems. A cloud provider can install and operate application software in a cloud (e.g., the network such as the Internet) and cloud users can access the application software from one or more of the client computing systems. Generally, cloud users that have a cloud-based site in the cloud cannot solely manage a cloud infrastructure or platform where the application software runs. Thus, the server computing systems and organized data structures thereof can be shared resources, where each cloud user is given a certain amount of dedicated use of the shared resources. Each cloud user's cloud-based site can be given a virtual amount of dedicated space and bandwidth in the cloud. Cloud applications can be different from other applications in their scalability, which can be achieved by cloning tasks onto multiple virtual machines at run-time to meet changing work demand. Load balancers distribute the work over the set of virtual machines. This process is transparent to the cloud user, who sees only a single access point.

[0252] Cloud-based remote access can be coded to utilize a protocol, such as Hypertext Transfer Protocol ("HTTP"), to engage in a request and response cycle with an application on a client computing system such as a web-browser application resident on the client computing system. The cloud-based remote access can be accessed by a smartphone, a desktop computer, a tablet, or any other client computing systems, anytime and/or anywhere. The cloud-based remote access is coded to engage in 1) the request and response cycle from all web browser based applications, 3) the request and response cycle from a dedicated on-line server, 4) the request and response cycle directly between a native

application resident on a client device and the cloud-based remote access to another client computing system, and 5) combinations of these.

Computing Devices

[0253] FIG. **9** illustrates a block diagram of an embodiment of one or more computing devices that can be used in combination with the present disclosure.

[0254] The computing device may include one or more processors or processing units **620** to execute instructions, one or more memories **630-632** to store information, one or more data input components **660-663** to receive data input from a user of the computing device **600**, one or more modules that include the management module, a network interface communication circuit **670** to establish a communication link to communicate with other computing devices external to the computing device, one or more sensors where an output from the sensors is used for sensing a specific triggering condition and then correspondingly generating one or more preprogrammed actions, a display screen **691** to display at least some of the information stored in the one or more memories **630-632** and other components. Note, portions of this design implemented in software **644**, **645**, **646** may be stored in the one or more memories **630-632** and be executed by the one or more processors **620**. The processing unit **620** may have one or more processing cores, which couples to a system bus **621** that couples various system components including the system memory **630**. The system bus **621** may be any of several types of bus structures selected from a memory bus, an interconnect fabric, a peripheral bus, and a local bus using any of a variety of bus architectures.

[0255] Computing device **602** typically includes a variety of computing machine-readable media. Machine-readable media can be any available media that can be accessed by computing device **602** and includes both volatile and non-volatile media, and removable and non-removable media. By way of example, and not limitation, computing machine-readable media use includes storage of information, such as computer-readable instructions, data structures, other executable software, or other data. Computer-storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other tangible medium which can be used to store the desired information, and which can be accessed by the computing device **602**. Transitory media such as wireless channels are not included in the machine-readable media. Machine-readable media typically embody computer readable instructions, data structures, and other executable software.

[0256] In an example, a volatile memory drive **641** is illustrated for storing portions of the operating system **644**, application programs **645**, other executable software **646**, and program data **647**.

[0257] A user may enter commands and information into the computing device **602** through input devices such as a keyboard, touchscreen, or software or hardware input buttons **662**, a microphone **663**, a pointing device and/or scrolling input component, such as a mouse, trackball, or touch pad **661**. The microphone **663** can cooperate with speech recognition software. These and other input devices are often connected to the processing unit **620** through a user input interface **660** that is coupled to the system bus **621** but can be connected by other interface and bus structures, such as a lighting port, game port, or a universal serial bus (USB). A display monitor **691** or other type of display screen device is also connected to the system bus **621** via an interface, such as a display interface **690**. In addition to the monitor **691**, computing devices may also include other peripheral output devices such as speakers **697**, a vibration device **699**, and other output devices, which may be connected through an output peripheral interface **695**.

[0258] The computing device **602** can operate in a networked environment using logical connections to one or more remote computers/client devices, such as a remote computing system **680**. The remote computing system **680** can a personal computer, a mobile computing device, a server, a router, a network PC, a peer device, or other common network node, and typically includes many or all of the elements described above relative to the computing device **602**. The logical connections can include a personal area network (PAN) **672** (e.g., Bluetooth®), a local area network (LAN) **671** (e.g., Wi-Fi), and a wide area network (WAN) **673** (e.g., cellular network). Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets, and the Internet. A browser application and/or one or more local apps may be resident on the computing device and stored in the memory.

[0259] When used in a LAN networking environment, the computing device **602** is connected to the LAN **671** through a network interface **670**, which can be, for example, a Bluetooth® or Wi-Fi adapter. When used in a WAN networking environment (e.g., Internet), the computing device **602** typically includes some means for establishing communications over the WAN **673**. With respect to mobile telecommunication technologies, for example, a radio interface, which can be internal or external, can be connected to the system bus **621** via the network interface **670**, or other appropriate mechanism. In a networked environment, other software depicted relative to the computing device **602**, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, remote application programs **685** as reside on remote computing device **680**. It will be appreciated that the network connections shown are examples and other means of establishing a communications link between the computing devices that may be used.

[0260] It should be noted that the present design can be carried out on a computing device such as that described with respect to FIG. **9**. However, the present design can be carried out on a server, a computing device devoted to message handling, or on a distributed system in which different portions of the present design are carried out on different parts of the distributed computing system.

[0261] Note, an application described herein includes but is not limited to software applications, mobile applications, and programs that are part of an operating system application. Some portions of this description are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical

quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like. These algorithms can be written in a number of different software programming languages such as C, C++, HTTP, Java, or other similar languages. Also, an algorithm can be implemented with lines of code in software, configured logic gates in software, or a combination of both. In an embodiment, the logic consists of electronic circuits that follow the rules of Boolean Logic, software that contain patterns of instructions, or any combination of both. A module may be implemented in hardware electronic components, software components, and a combination of both.

[0262] It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the above discussions, it is appreciated that throughout the description, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers, or other such information storage, transmission or display devices.

[0263] Generally, an application includes programs, routines, objects, widgets, plug-ins, and other similar structures that perform particular tasks or implement particular abstract data types. Those skilled in the art can implement the description and/or figures herein as computer-executable instructions, which can be embodied on any form of computing machine-readable media discussed herein.

[0264] Many functions performed by electronic hardware components can be duplicated by software emulation. Thus, a software program written to accomplish those same functions can emulate the functionality of the hardware components in input-output circuitry.

Email Cyber Threat Campaigns

[0265] Referring to FIG. **10**, a block diagram of an embodiment of an example chain of unusual behavior for the email(s) in connection with the rest of the network under analysis is shown. The user interface can display a graph **220** of an example chain of unusual behavior for the email(s) in connection with the rest of the network under analysis.

[0266] The network & email module can tie the alerts and events from the email realm to the alerts and events from the network realm.

[0267] The cyber-threat module cooperates with one or more machine learning models. The one or more machine learning models are trained and otherwise configured with mathematical algorithms to infer, for the cyber-threat analysis, 'what is possibly happening with the chain of distinct alerts and/or events, which came from the unusual pattern,' and then assign a threat risk associated with that distinct item of the chain of alerts and/or events forming the unusual pattern.

[0268] This is "a behavioral pattern analysis" of what are the unusual behaviors of the network/system/device/user/email under analysis by the cyber-threat module and the machine learning models. The cyber defense system uses unusual behavior deviating from the normal behavior and then builds a chain of unusual behavior and the causal links between the chains of unusual behavior to detect cyber threats. An example behavioral pattern analysis of what are the unusual behaviors may be as follows. The unusual pattern may be determined by filtering out what activities/events/alerts that fall within the window of what is the normal pattern of life for that network/system/device/user/email under analysis, and then the pattern of the behavior of the activities/events/alerts that are left, after the filtering, can be analyzed to determine whether that pattern is indicative of a behavior of a malicious actor—human, program, email, or other threat. The defense system can go back and pull in some of the filtered-out normal activities to help support or refute a possible hypothesis of whether that pattern is indicative of a behavior of a malicious actor. An example behavioral pattern included in the chain is shown in the graph over a time frame of, an example, seven days. The defense system detects a chain of anomalous behavior of unusual data transfers three times and unusual characteristics in emails in the monitored system three times which seem to have some causal link to the unusual data transfers. Likewise, twice unusual credentials attempted the unusual behavior of trying to gain access to sensitive areas or malicious IP addresses, and the user associated with the unusual credentials trying unusual behavior has a causal link to at least one of those three emails with unusual characteristics. When the behavioral pattern analysis of any individual behavior or of the chain as a group is believed to be indicative of a malicious threat, then a score of how confident the defense system in this assessment of is identifying whether the unusual pattern was caused by a malicious actor is created. Next, also assigned is a threat level parameter (e.g. score or probability) indicative of what level of threat does this malicious actor pose to the system. Lastly, the cyber-threat defense system is configurable in its user interface of the defense system on what type of automatic response actions, if any, the defense system may take when different types of cyber threats that are equal to or above a configurable level of threat posed by this malicious actor.

[0269] The cyber-threat module may chain the individual alerts and events that form the unusual pattern into a distinct item for cyber-threat analysis of that chain of distinct alerts and/or events. The cyber-threat module may reference the one or more machine learning models trained on e-mail threats to identify similar characteristics from the individual alerts and/or events forming the distinct item made up of the chain of alerts and/or events forming the unusual pattern.

[0270] One or more machine learning models may also be trained on characteristics and aspects of all manner of types of cyber threats to analyze the threat risk associated with the chain/cluster of alerts and/or events forming the unusual pattern. The machine learning technology, using advanced mathematics, can detect previously unidentified threats, without rules, and automatically defend networks.

[0271] The models may perform by the threat detection through a probabilistic change in normal behavior through the application of an unsupervised Bayesian probabilistic mathematical model to detect behavioral change in computers and computer networks. The Bayesian probabilistic

approach can determine periodicity in multiple time series data and identify changes across single and multiple time series data for the purpose of anomalous behavior detection. From the email and network raw sources of data, a large number of metrics can be derived each producing time series data for the given metric.

[0272] The detectors in the cyber-threat module including its network module and email module components can be discrete mathematical models that implement a specific mathematical method against different sets of variables with the target. Thus, each model is specifically targeted on the pattern of life of alerts and/or events coming from, for example, i) that cyber security analysis tool, ii) analyzing various aspects of the emails, iii) coming from specific devices and/or users within a system, etc.

[0273] At its core, the cyber-threat defense system mathematically characterizes what constitutes 'normal' behavior based on the analysis of a large number/set of different measures of a device's network behavior. The cyber-threat defense system can build a sophisticated 'pattern of life'—that understands what represents normality for every person, device, email activity, and network activity in the system being protected by the cyber-threat defense system.

[0274] As discussed, each machine learning model may be trained on specific aspects of the normal pattern of life for the system such as devices, users, network traffic flow, outputs from one or more cyber security analysis tools analyzing the system, email contact associations for each user, email characteristics, etc. The one or more machine learning models may use at least unsupervised learning algorithms to establish what is the normal pattern of life for the system. The machine learning models can train on both i) the historical normal distribution of alerts and events for that system as well as ii) factored in is a normal distribution information from similar peer systems to establish the normal pattern of life of the behavior of alerts and/or events for that system. Another set of machine learning models train on characteristics of emails and the activities and behavior of its email users to establish a normal for these.

[0275] Note, when the models leverage at least two different approaches to detecting anomalies: e.g. comparing each system's behavior to its own history, and comparing that system to its peers' history and/or e.g. comparing an email to both characteristics of emails and the activities and behavior of its email users, this multiple source comparison allows the models to avoid learning existing bad behavior as 'a normal' because compromised devices/users/components/emails will exhibit behavior different to their immediate peers.

[0276] In addition, the one or more machine learning models can use the comparison of i) the normal pattern of life for that system corresponding to the historical normal distribution of alerts and events for that system mapped out in the same multiple dimension space to ii) the current chain of individual alerts and events behavior under analysis. This comparison can yield detection of the one or more unusual patterns of behavior within the plotted individual alerts and/or events, which allows the detection of previously unidentified cyber threats compared to finding cyber threats with merely predefined descriptive objects and/or signatures. Thus, increasingly intelligent malicious cyber threats that try to pick and choose when they take their actions in order to generate low level alerts and event will still be detected, even though they have not yet been identified by other methods of cyber analysis. These intelligent malicious cyber threats can include malware, spyware, key loggers, malicious links in an email, malicious attachments in an email, etc. as well as nefarious internal information technology staff who know intimately how to not set off any high level alerts or events.

[0277] In essence, the plotting and comparison is a way to filter out what is normal for that system and then be able to focus the analysis on what is abnormal or unusual for that system. Then, for each hypothesis of what could be happening with the chain of unusual events and/or alerts, the gatherer module may gather additional metrics from the data store including the pool of metrics originally considered 'normal behavior' to support or refute each possible hypothesis of what could be happening with this chain of unusual behavior under analysis.

[0278] Note, each of the individual alerts and/or events in a chain of alerts and/or events that form the unusual pattern can indicate subtle abnormal behavior; and thus, each alert and/or event can have a low threat risk associated with that individual alert and/or event. However, when analyzed as a distinct chain/grouping of alerts and/or events behavior forming the chain of unusual pattern by the one or more machine learning models, then that distinct chain of alerts and/or events can be determined to now have a much higher threat risk than any of the individual alerts and/or events in the chain.

[0279] Note, in addition, today's cyberattacks can be of such severity and speed that a human response cannot happen quickly enough. Thanks to these self-learning advances, it is now possible for a machine to uncover these emerging threats and deploy appropriate, real-time responses to fight back against the most serious cyber threats.

[0280] The threat detection system has the ability to self-learn and detect normality in order to spot true anomalies, allowing organizations of all sizes to understand the behavior of users and machines on their networks at both an individual and group level. Monitoring behaviors, rather than using predefined descriptive objects and/or signatures, means that more attacks can be spotted ahead of time and extremely subtle indicators of wrongdoing can be detected. Unlike traditional legacy defenses, a specific attack type or new malware does not have to have been seen first before it can be detected. A behavioral defense approach mathematically models both machine, email, and human activity behaviorally, at and after the point of compromise, in order to predict and catch today's increasingly sophisticated cyber-attack vectors. It is thus possible to computationally establish what is normal, in order to then detect what is abnormal. In addition, the machine learning constantly revisits assumptions about behavior, using probabilistic mathematics. The cyber-threat defense system's unsupervised machine learning methods do not require training data with pre-defined labels. Instead, they are able to identify key patterns and trends in the data, without the need for human input.

[0281] The cyber-threat defense system **100** may use at least three separate machine-learning models. Each machine-learning model may be trained on specific aspects of the normal pattern of life for the system such as devices, users, network traffic flow, outputs from one or more cyber security analysis tools analyzing the system, etc. One or more machine learning models may also be trained on

characteristics and aspects of all manner of types of cyber threats. One or more machine learning models may also be trained on characteristics of emails themselves. In an embodiment, the one or more models may be trained on specific aspects of these broader concepts. For example, the models may be specifically trained on associations, attachments, compliances, data loss & transfers, general, metadata, hygiene, links, proximity, spoof, type, validation, and other anomalies.

[0282] In addition, the one or more machine learning models can be self-learning using unsupervised learning algorithms. For example, a set of the one or more machine learning models can be trained on the normal behavior of users and their emails use data from the probes to train on; and therefore, regularly update what a base line for the normal behavior is. This autonomous, self-learning defense system protects against malicious activity in the email domain—whether the malicious activity is from any of i) standard threat actors from email, such as phishing and malware emails, and ii) insider threat from users, which does not rely solely on pre-recognized, arbitrary ideas of malicious email domain activity but instead autonomously contextualizes each communication to assess its anomaly compared to standard behavior of the user and organization.

[0283] Referring to FIG. **11**, a block diagram of an embodiment of an email similarity scoring module configured to cooperate with the one or more mathematical models in order to compare an incoming email, based on a semantic similarity of multiple aspects of the email to a cluster of different metrics derived from known bad emails to derive a similarity score between an email under analysis and the cluster of different metrics derived from known bad emails is shown.

[0284] The similarity to known bad emails detector looks for a cluster of different metrics derived from known bad emails (content, language-usage, subjects, sentence construction etc.), which is then compared to all inbound and/or outbound emails to derive a similarity. The known bad emails are sourced from a data set of previously observed emails by the system which were deemed clearly malicious or contained known bad content as identified by internal and external threat intelligence. A comparison is made on the multiple aspects of the email under analysis to those same multiple metrics of known bad emails. The mathematical models use an algorithm to weigh different metrics, emphasize or throw out outliers, and establish an overall similarity between multiple metrics of an email under analysis and multiple metrics of each of the known bad emails. A threshold is determined to declare the two emails are similar enough. Additional factors, such as the layout of the email changing from its historical norm, the email going to a rare grouping of multiple recipients, a presence of an image tracking link that is purposely being concealed, etc., can all be included in the analysis to lower the threshold to declare an email under analysis as potentially suspicious and needing an autonomous action to contain the email.

[0285] Note, the cyber-threat defense system **100** quickly referencing a similarity score between compared emails also allows the system to find other similar emails when the cyber threat module determines a particular email is indeed malicious in nature. Thus, the models can fairly quickly analyze similarity of emails based on headers, subject, body, links, attachments, etc. Then, even if the e-mails have different senders and recipients and possibly different subject lines, the other characteristics of the email may be mathematically deduced to be very similar. Once the email similarity scoring module finds one email determined to be malicious by the cyber threat module, then the email similarity scoring module can go out and find other similar e-mails that are also likely malicious, even if they have different aspects such as different senders and recipients. Although some changes may differ between the headers, subject line, body, links, attachments, sender, domain, etc., statistically enough overlapping similarity exists to think that the other similar e-mails are likely also suspicious e-mails. These emails are fed into the dataset to expand the awareness of "known bad."

[0286] Note, individual metrics derived from a known bad email can also often mutate and/or slightly change. Thus, the cluster of different metrics is analyzed together versus merely matching a signature of a known bad email with its fixed characteristics. The similarity to known bad emails detector looks for a cluster of different metrics derived from known bad emails (content, language-usage, subjects, sentence construction etc.), which is then compared to all inbound and/or outbound emails to derive an overall similarity for the group of metrics being compared. The email similarity scoring module also records metrics and regularly updates metrics about known bad email communications it has observed, which are then modeled and compared with new emails as they arrive to identify any similarity, detecting attacks which have slightly changed from the "known" methodology or from new unique senders.

[0287] In addition, establishing an email similarity to aspects/characteristics of known bad emails allows 1) the cyber threat module to identify previously unknown malicious emails as well as 2) the autonomous response module to know what appropriate autonomous response actions to take against that type of attack based on successful responses to similar bad emails, from previously unseen sender emails coming to a user's inbox. The combining analysis and modeling of known bad emails, via the email similarity scoring module, and known changes in user behavior, via the layout change predictor module, as well as analysis via machine learning, allow the ability to reliably detect unknown, previously unseen emails from senders as malicious.

[0288] In addition, updating of the models occurs. The email similarity scoring module also records metrics about known bad email communications it has observed, which are then modeled, added to the collection of known bad emails, and compared with new emails as they arrive to identify any similar email including those that i) have slightly changed from the 'known' methodology and/or ii) come from new unique senders.

[0289] In an exemplary embodiment as shown in FIG. **11**, the CSA **100** may determine if an email body similarity to known bad emails score is greater than a defined threshold. If so, the cyber security system may take appropriate action. The CSA **100** may determine if email links similarity to known bad emails score is greater than a defined threshold. If so, the cyber security system takes appropriate action. The CSA **100** may determine if email header similarity to known bad emails score is greater than a defined threshold. If so, the cyber security system takes appropriate action.

[0290] Referring to FIG. **12**, a block diagram of an embodiment of a mass email association detector configured to determine a similarity between two or more highly similar emails being i) sent from or ii) received by a collection of

two or more individual users in the email domain in a substantially simultaneous time frame, where one or more mathematical models are used to determine similarity weighing in order to derive a similarity score between compared emails is shown. The mathematical models configured to determine similarity discussed above can determine if two or more emails are highly similar via the score from the compared metrics.

[0291] The mass email association module determines what is a likelihood for two or more highly similar emails that are being i) sent from or ii) received by a collection of individuals in the email domain under analysis in a substantially simultaneous time period. Note, the substantially simultaneous time period can be equal to or less than a ten second difference in any of i) a time sent for each of the similar emails under analysis, and ii) a time received for each of the similar emails under analysis. The mass email association module determines the likelihood based on at least i) historical patterns of communication between those individuals, ii) how rare it is for that particular collection of individuals to all send and/or receive this highly similar email in roughly a same time frame, or iii) how rare it is for individuals identified by this system as having a similar pattern of activity and communication to those receiving the communication to send and/or receive this highly similar email in roughly a same time frame. A low likelihood is indicative that the communicator of the similar email being sent out in mass had no prior association with those individuals; and is therefore, more likely to be malicious in intent.

[0292] The mass email association module creates a map of all contacts that are usually addressed in the same (or similar) emails, for every inbound and outbound email in this email domain. This map is then used to derive a probability likelihood that the collection of individuals, e.g. two users, would be included in the same email. The email association scoring module with its mapping and algorithms allows detection of unusual combinations of recipients, whether they are unintentionally added to the email or whether this is indicative of a supply-chain compromise or an attacker attempting to add legitimacy to an attack.

[0293] The cyber threat defense system **100** for email derives a wide range of metadata from observed email communications which it analyzes with one or more machine learning models to form a 'pattern-of-life' of user activity and email activity for the email system of a given organization. This pattern-of-life recognizes and maps associations between users to generate a probabilistic likelihood that the two or more users would be included in the same communication, decreasing false positive rates, and identifying intelligent attacks. This baseline 'normal' for the organization includes a fingerprint of email communications for all internal and external senders seen which is compared with any new communications to detect subtle behavioral shifts that may indicate compromise.

[0294] One or more machine learning models are trained to gain an understanding of a normal behavior of email activity and user activity associated with email domain. For example, the models train on content a user of the network views and/or sites frequented inside and outside of the network as well as checks e-mail history to see if it is probable that this email user would be receiving this particular e-mail under analysis. The models train on e-mail usage pattern of life, content style, contacts, and group associations of each e-mail users in that system. The models cooperating with the module can then determine what is the likelihood that this e-mail under analysis falls outside of that normal behavior of email activity and user activity for that e-mail user. The module's analysis is married/combined with an output from one or more machine learning models trained on gaining an understanding of all of the characteristics of each e-mail itself and its related data, classify properties of the e-mail, and what is the likelihood the e-mail under analysis falls outside of being a normal benign email. Combining both analyses can allow the mass email association module to determine a likelihood of multiple emails that are highly similar in nature being i) sent to or ii) received by the collection of individuals targeted by that mass mailing of highly similar emails currently under analysis, all sent out at about the same time. The e-mail may be highly similar in nature when a comparison of the emails indicate that they have, for example, similar content, similar subject line, similar sender and/or domain, etc.; and thus, share many similar characteristics.

[0295] Note, the cyber threat defense system can also factor associations and anomalies that do not correspond to associations for that email account and/or user. The cyber threat defense system works out whom the email account and/or the user know such as in their contacts, whom the email account and/or the user communicate with, and other factors like e-mail addresses from a same organization. Thus, the cyber threat defense system can work out, in essence, how many degrees of separation exist between the sender of an e-mail and the email account and/or user. The cyber threat defense system can then use these associations as a factor in the determination of the likelihood that a substantially similar email would be sent to or be received by that set of e-mail recipients. The module factors into the likelihood determination factors such as historical trends for that network for associations, entries in the users list of contacts, how close in time the emails are sent, similarity in content and other aspects, how unusual that collection of individual are being grouped together in this email, etc.

[0296] In an exemplary embodiment as shown in FIG. **12**, the CSA **100** may determine if the four criteria: (i) an email characteristics meet specified criteria, (ii) an email reply characteristics meet specified criteria, (iii) the sender's popularity is greater than a defined threshold, and (iv) the recipients' association anomaly score is greater than a defined threshold. If criterion (i) is true, criterion (ii) is false, if criterion (iii) is false, and criterion (iv) is true, then the cyber security system takes appropriate action.

[0297] Referring to FIG. **13**, a block diagram of an embodiment of an email module and network module cooperating to supply a particular user's network activity tied to their email activity is shown. The relationships between an embodiment of the email module and how the relationship between email addresses may be visually presented on the user interface module **150**. Recipients, senders, contact lists for each user may be diagrammed to see how close of a relationship exists; and thus, a factor of how likely or unusual this recipient is to receive an email from this sender; and vice versa.

[0298] The network module and its machine learning models as well as the email module and its machine learning models are utilized to determine potentially unusual network activity that provides an additional input of information into the cyber-threat module to determine the threat risk param-

eter. A particular user's network activity can be tied to their email activity because the network module observes network activity, and the cyber-threat module receives the network module observations to draw that into an understanding of this particular user's email activity to make an appraisal of potential email threats with a resulting threat risk parameter tailored for different users in the e-mail system.

[0299] Sender Interactions: A first pane of the user interface graphically represents an example of an email interaction observed by the email module for the sender email address. The sender node is the central node, and the recipient for the specific message selected is indicated by a larger connected node.

[0300] Recipient Interactions: A second pane of the user interface graphically represents an example of all of the email interactions observed by the email module for the recipient email address. The recipient node is the central node, and the sender for the specific message selected is indicated by a larger connected node.

[0301] The email module keeps track of whether a domain is internal or external in relation to the email application's domain that it is monitoring. Therefore, for external recipients/senders, others from their organization or domain will also appear as external.

[0302] Referring to FIG. **14**, a block diagram of an example Cyber Security Appliance (CSA) configured to detect, analyze, and act against email cyber threat campaigns is shown. CSA **100** may comprise an email data store where relevant email data, metadata, comprehensive data logs, and the emails themselves may be stored, and an AI model(s) store where modules that are common between different modules may be stored. While shown internal to CSA **100**, in certain embodiments, it may be located elsewhere on the network, like, for example, a server, disc array, etc.

[0303] CSA **100** may further comprise a User Interface Module (UIM) coupled to the email data store and the AI model(s) data store, and that may be configured to be coupled to Keyboard/Video/Mouse (KVM) hardware and a video display. The UIM may enable a human cyber security professional to see all of the relevant data in one place in various formats and to input commands via the KVM in response to potential threats.

[0304] The UIM can graphically display logic, data, and other details that the cyber-threat defense system goes through. The cyber-threat module, in cooperation with machine learning models, analyzes these metrics in order to develop a rich pattern of life for the user activity and email activity in that email system. This allows the cyber-threat module, in cooperation with the email module (discussed below), to spot unusual anomalous emails and/or behavior that have bypassed/gotten past the existing email gateway defenses.

[0305] In some embodiments, the UIM can display a graphic showing all of the interactions of an email of interest as described in conjunction with FIG. **13** above. In other embodiments, the UIM may generate standard or custom reports for display or printout. In various embodiments, there may be a graphical database with the necessary data to make clear video displays that display an email cyber threat campaign in a single plane of analysis.

[0306] In various other embodiments, the CSA **100** may relate email relationships in a Graphical Database. The CSA **100** aims to render an email network as a series of nodes in a graph database, where the communications between those nodes are the edges. The edges may distinguish between types of communication and will be weighted to represent the strength of the communication. The email relationships are mapped out as the nodes in the graph database and then updated on a dynamic basis over time as the email system continues to operate. The UIM is further configured to display on a display screen trends for one or more ongoing email campaign detections, and the emails involved in the email campaign and the email accounts that have been targeted.

[0307] In certain embodiments, there may be another (trustworthy) cyber security appliance (not shown) coupled to the network and configured to communicate with CSA **100**. In such cases, the UIM is configured to coordinate with the other cyber security appliance in sharing data, communicating when an email campaign is detected, and assisting in formulating responses against detected email cyber threat campaigns.

[0308] In certain other embodiments, the UIM may display tracking of the various high-interest entities in the network. Examples of high-interest entities may be high-permission users, Software as a Service (SaaS) files, macro-enabled documents, etc. If necessary, the UIM may display a SaaS console for viewing by the cyber security professional. In many embodiments, all files in the system may be tracked.

[0309] This approach may comprise one or more risk profiles for an entity, such as a worker, derived from their activity tracked in the multiple different cyber protection domains. The risk profiles may incorporate the amount of exposure and protections involved in that exposure to external actors (e.g., other email users and SaaS users) and their permissions levels in, for example, the SaaS platform. These can also be looked at to increase that user's risk profile based on the combination of risks in the email domain and risks in the SaaS domain. Some embodiments can be used for cross-platform profiling to generate risk profiles. The example embodiments may be suggested for cross-platform profiling, which indicates users who are 'risky' based upon their exposure to, for example, external actors in emails or their permissions level within SaaS platforms.

[0310] Users who are exposed to many external actors in their email communications and have access to a wide range of files internally, or who have 'high' permission levels or are members of high-level groups can be profiled as 'risky' due to the impact their compromise by a cyber threat could have.

[0311] Users who are found to be high-permissions users in the SaaS environment can trigger higher anomalies when potentially malicious emails are detected due to the impact their compromise by a cyber threat could have.

[0312] Files accessed in the SaaS environment are checked against those recently seen in email communications entering the organizations. This is a classic attack vector for things like macro-enabled documents.

[0313] Emails can be checked for instances of SaaS services that are not covered by the CSA **100**, indicating potential 'shadow' accounts.

[0314] The riskiness of the location of a SaaS account may be evaluated based upon known-active email threat campaigns in their geographic location. Active cyber threat

campaigns can occur in geographic locations, and a simple presence in that geographic locations can increase the risk profile.

[0315] The UIM for a SaaS console can be the one plane to analyze and present these cross-platforming risk profiles. In one or more embodiments, the email CSA **100** may talk to the SaaS cyber security appliance, etc.

[0316] Additionally, files can be tracked on entry into the organization, emails can be checked against known SaaS accounts to detect unauthorized SaaS platform usage, and user location as retrieved from the SaaS platform can be compared to active malicious email campaigns to increase alert scores. These implementations are not limited to one platform (e.g., the email platform could be Gmail and the SaaS platform JumpCloud), and AWS permissions could also be examined, and the use of a Dropbox account authorized by the company or not, etc., are all example factors looked at in the risk profile.

[0317] The UIM may accept feedback from a user who may confirm or reject inferences from the various modules and AI models in the system. The CSA **100** may take this input from a cyber security professional and feed it back to the training sets for the various AI and machine learning models in the system.

[0318] CSA **100** may further comprise an Email Module (EM) that may be coupled to the email data store, the AI model(s) data store, and the UIM. The EM may monitor all of the routine email functions of an email application (e.g., receiving and displaying emails and attachments, composing emails, adding attachments, sending responses, etc.). The email module may provide comprehensive email logs for every email observed. These logs can be filtered with complex logical queries, and each email can be interrogated on a vast number of metrics in the email information stored in the data store. Some example email characteristics that can be stored and analyzed are:

[0319] Email direction: Outbound emails and inbound emails.

[0320] Send Time: The send time is the time and date the email was originally sent according to the message metadata.

[0321] Links: Every web link present in an email has its own properties. Links to websites are extracted from the body of the email. Various attributes are extracted, including, but not limited to, the position in the text, the domain, the frequency of appearance of the domain in other emails and how it relates to the anomaly score of those emails, how well that domain fits into the normal pattern of life of the intended recipient of the email, their deduced peer group, and their organization.

[0322] Recipient: The recipient of the email. If the email was addressed to multiple recipients, these can each be viewed as the "Recipients." The known identifying properties of the email recipient, including how well known the recipient was to the sender, descriptors of the volume of mail, and how the email has changed over time, to what extent the recipient's email domain is interacted with inside the network

[0323] Subject: The email subject line.

[0324] Attachment: Every attachment associated with the message will appear in the user interface here as individual entries, with each entry interrogatable against both displayed and advanced metrics. These include, but are not limited to, the attachment file name, detected file types, descriptors of the likelihood of the recipient receiving such a file, descriptors of the distribution of files such as these in all emails against the varying anomaly score of those emails.

[0325] Headers: Email headers are lines of metadata that accompany each message, providing key information such as sender, recipient, and message content type, for example.

[0326] The email module's underlying structure may enable the detection of anomalous flows of communication between and from the business (different cyber domains edges) and for the identification of highly connected nodes (i.e., at-risk users). Normal behavior can be modelled using a connectivity web. The EM can employ different types of Machine Learning (ML) approaches on the dataset. The ML models may be present in the EM module or reside in the AI Model(s) Store as a matter of design choice.

[0327] During the analysis, the email module can reference the one or more machine learning models that are self-learning models trained on the normal behavior of email activity and user activity associated with an email system. This can include various email policies and rules that are set for this email system. The cyber threat module may also reference the models that are trained on the normal characteristics of the email itself. The cyber threat module (discussed below) can apply these various trained machine learning models to data, including metrics, alerts, events, metadata from the network module and the email module, etc. In addition, a set of AI models may be responsible for learning the normal 'pattern of life' for internal and external address identities in connection with the rest of the network for each email user. This enables the system to neutralize malicious emails which deviate from the normal "pattern of life" for a given address identity for that user in relation to its past, its peer group, and the wider organization.

[0328] The EM may have at least a first email probe to inspect an email at the point it transits through the email application, such as Office 365, and extracts hundreds of data points from the raw email content and historical email behavior of the sender and the recipient. These metrics are combined with the pattern of life data of the intended recipient or sender, sourced from the data store. The combined set of the metrics is passed through machine learning models to produce a single anomaly score of the email, and various combinations of metrics will attempt to generate notifications which will help define the 'type' of email. None, some, or all of the AI models used by the EM may reside internal to the EM or be located in the AI Model(s) store.

[0329] CSA **100** may further comprise a Cyber Threat Module (CTM) that may be coupled to the email data store, the AI model(s) data store, the UIM, and the EM. The email module detects emails whose content is not in keeping with the normal pattern of content as received by this particular recipient and sends the data, metadata, email, and log information to the CTM for analysis. As discussed previously, in conjunction with FIG. **4**, an example analysis may be as follows:

[0330] (i) To what level has the sender of this email previously communicated with individuals within the receiving organization?

[0331] (ii) How closely are the recipients of this mail related to those individuals who have previously communicated with the sender?

[0332] (iii) Is the content of this email consistent with other emails that the intended recipient sends or receives?

[0333] (iv) If any links or attachments present in the email were to be clicked or opened by the intended recipient, would this constitute anomalous activity for that individual's normal network behavior?

[0334] (v) Are the email properties consistent with this particular user's recent network activities?

[0335] Thus, the cyber-threat module can also reference the machine learning models trained on an email itself and its related data to determine if an email or a set of emails under analysis have potentially malicious characteristics. One or more of the Artificial Intelligence (AI) models may trained on one or more of the group of parameters consisting of: (i) email metrics, (ii) suspicious characteristics of emails, (iii) malicious links and attachments, and (iv) a database of known malicious actors, and (v) any combination of these parameters, in order to cluster emails with similar parameters The cyber-threat module can also factor these parameters and email characteristics analysis into its determination of the threat risk parameter.

[0336] The email module can coordinate with an AI model trained to model the email behavior in the system to retrospectively process an email application's metadata, such as Office 365 metadata, to gain an intimate knowledge of each of their users, email addresses, correspondents, and routine operations. The power of the cyber-threat module lies in leveraging this unique understanding of day-to-day user email behavior, of each of the email users, in relation to their past, their peer group, and the wider organization. Armed with the knowledge of what is 'normal' for a specific organization and specific individual, rather than what fits a predefined template of malicious communications, the cyber-threat module can identify subtle, sophisticated email campaigns which mimic benign communications and locate threats concealed as everyday activity (see, e.g., FIG. **4** for a visual representation of an email address' association data).

[0337] CSA **100** may further comprise an AI Campaign Detector Module (ACDM) that may be coupled to the email data store, the AI Model(s) data store, the UIM, the EM, and the CTM. A trigger for an email to be considered for the ACDM may be that it has been considered above the threshold for anomaly by an existing AI-based email classifier (e.g., the CTM or other module or AI model in the system) and, therefore, a "significant action" may have already been taken against it.

[0338] The system can look at the rarity across all emails in the campaign, what is the normal score the CTM has given them, and feed that all into a meta classifier for building detection of an email campaign that gets bits from other AI classifiers. Campaign detection is performed through the ACDM, which is trained to perform a meta-analysis of the output from existing AI-based email classifiers.

[0339] Emails that are modeled as similar are compared with popularity metrics such as a counter (e.g., to avoid considering everyone receiving the same newsletter a "campaign"). Then, those that are highly similar may be compiled into a campaign. This may also be performed retroactively to identify the start of the campaign—any additional emails found may be upgraded in severity of autonomous action taken against that email to match the whole campaign.

[0340] To identify campaigns, we may need to understand email addresses that are closely linked together and those that are externally facing (or public-facing). Externally facing addresses may be expected to receive large quantities of benign mail of low quality, which could be detected as false positives for campaign detection. Internally facing (or private-facing) mailboxes tend to be more at risk because someone would have to know that email address to target it. Such private facing boxes may belong to executives and other high-permission users whose accounts could be quite damaging if compromised. Also, the target may be more at risk due to their position in the organization. This may be done by a meta-analysis where the higher the anomaly score, the more likely the email account was not publicly facing (e.g., as any attacks that did come in were highly targeted) versus low-quality spam to public email addresses (like, for example, "sales@company.com" or "info@organization.org").

[0341] Determining the purpose of a mailbox may be determined by the ACDM using meta-scoring. Many corporate inboxes receive large quantities of email from varied and unpredictable external servers. For some inboxes, it is both expected and intended because they have public-facing addresses, and most of the mail they receive from external senders is usually characterized by benign intent. For others, such as those belonging to company executives, messages sent by external senders are much more likely to be malicious, irritating, or both. The autonomous detector wants to know in a corporate environment which email inboxes get hammered with the most malicious emails because, potentially due to their public exposure and they can represent key users.

[0342] The ACDM is able to analyze every inbox in an environment by analyzing existing threat scores, assigning the scores to categories, and logging them over time. Counts are fed into a probability function that maintains a score for each inbox, thereby estimating the likelihood it should receive unsolicited mail. Following a period of learning in an example live environment, the ACDM was able to identify the public-facing mailboxes with very few false positives. By incorporating the ACDM into the existing autonomous response system, better-tailored actions could be taken against unsolicited emails.

[0343] Autonomous detection of an intended function of a corporate inbox and whether the corporate inbox under analysis that is receiving a higher number of malicious emails than the typical corporate inbox is associated with a low priority intended recipient (account or user) or a high priority intended recipient. When the intended recipient is a high priority intended recipient, such as a CEO, then the autonomous response should increase its level of stringency in the actions it takes against those malicious emails. (e.g., not pass attached documents and/or links without a quarantine or stripping. When the intended recipient is a low priority intended recipient, such as the published corporate inbox info@companyname.com, then the autonomous response can maintain the level of stringency in the actions it takes against those malicious emails. Inboxes for low priority intended recipients typically either have extra stringency on how emails are handled and/or have limited exposure to other key portions of a network. Autonomous detection can also look at an anomalous level/severity level of a suspicious emails, such as highly targeted phishing email, to determine a priority level of an intended recipient.

[0344] Mapping the relationship between contacts may also be necessary. By utilizing a graphical database, close relations between internal and external contacts can be

identified as well as close relationships between internal contacts as recipients. This may be useful information when detecting a campaign as it can be posited that lots of internal email addresses which are not closely connected suggest an email blast campaign more than a collection of internal addresses that are regularly clustered together as recipients. Thus, email relationships in a graphical database detection of anomalous flows of communication within and outside the business (low weighted edges) and the identification of highly connected nodes (i.e., at-risk users). Similarly, the more distant the relationship between external contacts and internal recipients, the more likely the communication is not desirable and may represent a campaign.

[0345] The ACDM may be trained to compare all clustered emails in the last fixed timeframe selectable by the user (e.g., a week, two weeks, etc.), looking for similarities and then looking for specific hallmarks. This information can be compared to a database of known malicious actors, sources on the internet, and other information. Many organizations get thousands of emails per day. Larger organizations may receive hundreds of thousands of emails a day. Grouping emails and providing this automated analysis can be very helpful in identifying and understanding email threats. The risk may be estimated for the campaign utilizing existing email metrics (e.g., inducement metrics—is this spam, or is it highly targeted phishing?).

[0346] The ACDM may be trained to examine multiple factors on these emails to also identify who is the threat actor engaging in the campaign as a targeted attack on the organization. The ACDM may be configured to try and understand the purpose of a campaign rather than merely stopping the emails of the campaign attack. The ACDM may be trained to analyze factors such as the wording of the email, malicious links and/or attachments removed, and other metadata analysis to make a geographical location estimation (e.g., same host or domain), as well as identify other characteristics shared by the emails of the campaign. This information can be compared to a database of known malicious actors, sources on the internet, and other information.

[0347] The ACDM may also use internal classifications and machine learning systems to assist in determining who is targeting an organization, why they are targeting the organization, and where the email campaign is coming from. Campaigns may be ranked within an environment by the level of risk they pose to the recipients. For example, an inducement classifier can be trained to evaluate the inducement across all of the emails in the campaign.

[0348] The ACDM may be able to assess the popularity and/or the number of emails associated with a campaign similarity, and then may classify the interestingness of the email campaign, and then may be able to identify who is targeting the organization, why they are targeting the organization, and where the email campaign is coming from.

[0349] The ACDM may determine whether or not the emails under consideration are interesting, and then the system may format all that information into a human-readable presentation (e.g., display screen and/or printed report presented to the end-user) through the UIM. The ACDM may detect email campaigns, rank these email campaigns, and then explain to the end-user why the organization is being targeted, who is targeting them, and what they are trying to achieve.

[0350] The CSA **100** may relate email relationships in a graph database. The CSA **100** may aim to render an email network as a series of nodes in a graph database, where the communications between those nodes are the edges. The edges will distinguish between types of communication and will be weighted to represent a strength of the communication. The email relationships are mapped out as the nodes in the graph database and then updated on a dynamic basis over time as the email system continues to operate. The ACDM may also use email relationships in a graph database detection of anomalous flows of communication between and from the business (low weighted edges) and for the identification of highly connected nodes (i.e., at-risk users).

[0351] CSA **100** may further comprise an Autonomous Response Module (ARM) that may be coupled to the email data store, the AI model(s) data store, the UIM, the EM, the CTM, and the ACDM. Autonomous action may be taken against email cyber threat campaigns by the ARM, which may act in conjunction with the other modules. A detected email cyber threat campaign may be compared to previously detected campaigns, and the detected campaign may be analyzed to understand its origins and purposes. The analysis may be performed by the ACDM in conjunction with other modules and one or more AI models.

[0352] Autonomous action may also be taken against the individual email cyber threats that make up a campaign. These actions may be taken by the ARM or other modules. In particular, the autonomous response module is configured to take an autonomous action against the email, where the email acted against had previously been examined and not stopped or acted upon but now is deemed part of an email campaign. In response, a malicious attachment may be stripped from an email threat to neutralize a macro-enabled document or one containing a malicious link and/or an email header may be unspoofed, allowing the recipient to see where the email is truly from instead of it being concealed behind a false header.

[0353] A malicious link in an email threat may be autonomously acted upon in a number of different ways. A link may be deleted to prevent a user from clicking on it. A link may be locked (the URL of the link may be replaced such that a click of that link will first divert the user via an alternative destination). A link may be double-locked. In this case, the URL may be replaced with a redirected email link. If the link is clicked, the user will be presented with a notification to that user that they are not permitted to access the original destination of the link. The user will be unable to follow the link to the original source, but their intent to follow the link may be recorded for a subsequent follow-up with the user. Other autonomous actions may include holding back a message and not delivering it to an email mailbox or moving the email threat to a junk mailbox.

[0354] It should be noted that the various AI models employed in the various modules of CSA **100** may reside either in the AI model(s) data store or internally to the various modules as a matter of design choice.

[0355] Referring to FIGS. **15**A, **15**B, and **15**C, a flowchart diagram depicting a process of the operation of CSA configured to detect, analyze, and act against email cyber threat campaigns is shown. Process **1500** may begin by processing all incoming mails (block **1510**). An Email Module (EM) may process the data, and logs may be generated by the email module. The logs may comprise data and metadata from the incoming emails (block **1520**). The likelihood and

level of an email cyber threat may be determined by one or more AI models working in conjunction with an EM, a Cyber Threat Module (CTM), or another module (block **1530**).

[0356] The likelihood of an email cyber threat campaign may be determined (block **1540**). An AI Campaign Detector Module (ACDM) may make the determination. An email may be detected as a threat by an EM, a CTM, or another module before being presented to an ACDM for analysis. The email threats may be scored according to a plurality of indices stored in the logs from an EM, in an email data store, or elsewhere in a system (block **1541**). The scoring may be done by one or more AI models and may be used to group similar emails into clusters (block **1542**). The scoring may also be used to determine the fluctuations of the incoming email threats by examining their chronological sequence (block **1543**). Emails that are part of a campaign may be similar in various ways and may score similarly. This provides a self-correction aspect in that if there is a large variation in the scoring of the individual email threats (fluctuations) over a distribution, then they are likely not part of a campaign. Similarly, if the variation of the scoring is smooth (many similar scores), then the email threats are likely part of a campaign.

[0357] Each of the mailboxes in a system may be analyzed to determine its purpose (block **1544**) and if it is public-facing or private-facing (block **1545**). Various factors like the facing, the user associated with the mailbox, the permissions of the user, the patterns of use of both the user and the mailbox itself, etc., may be used to estimate the probability of unsolicited emails to each mailbox in the system (block **1546**). The number of threats to each mailbox is logged (block **1547**), and each mailbox is continuously scored based on the incoming threats (block **1548**).

[0358] Autonomous action may be taken against email cyber threat campaigns (block **1550**). An Autonomous Response Module (ARM) may act in conjunction with other modules like, for example, an EM, an CTM, an ACDM, etc. A detected email cyber threat campaign may be compared to previously detected campaigns (block **1551**), and the detected campaign may be analyzed to understand its origins and purposes (block **1552**). The analysis may be performed by an ACDM in conjunction with other modules and/or a variety of AI models.

[0359] Autonomous action may also be taken against the individual email cyber threats that make up a campaign (block **1560**). These actions may be taken by an ARM or other modules. A malicious attachment may be stripped from an email threat (block **1561**). This would be an appropriate way to handle a macro-enabled document or one containing a malicious link. An email header may be unspoofed (block **1562**). This allows the recipient to see where the email is truly from instead of it being concealed behind a false header.

[0360] A malicious link may be autonomously acted upon in a number of different ways. A link may simply be deleted to prevent a user from clicking on it (block **1563**). A link may be locked, and the URL of the link may be replaced such that a click of that link will first divert the user to an alternative destination (block **1564**). A link may be double-locked, and the URL may be replaced with a redirected email link (block **1565**). If the link is clicked, the user will be presented with a notification to that user that they are not permitted to access the original destination of the link. The user will be unable to follow the link to the original source, but their intent to follow the link may be recorded for a subsequent follow-up with the user. Other autonomous actions may include holding back a message and not delivering it to an email mailbox (block **1566**) and moving the email threat to a junk mailbox (block **1567**).

[0361] Reporting on detected cyber threat emails and email campaigns may be accomplished in a plurality of modes (block **1570**). A User Interface Module (UIM) may be used. One approach is to generate a report (block **1571**), while another is to display the data on a screen (block **1572**). This may aid a cyber security professional in understanding the situation both to facilitate action and to explain the results to non-professionals.

[0362] A graphical display of the system indicating the details of a cyber threat campaign may be generated (block **1573**). This may include providing a single plane of analysis (block **1574**). For example, a user interface for a SaaS console can be the one plane to analyze and present cross-platforming risk profiles. A graphical display can show all the pertinent information (including risk profiles) and provide profiles of the various entities involved like, for example, networks, systems, users, nodes, edges, etc., (block **1575**) and how they are related using a connectivity web (block **1576**).

[0363] The tracking of various entities may also be displayed. For example, high permission users may be tracked (block **1577**). These users are at higher risk because if they or their accounts are compromised, the resulting damage may be more severe. Various document types can be tracked. For example, SaaS documents and macro-enabled documents may be tracked because of the danger they pose if compromised (block **1578**). Further, all documents in an entire system or organization may be tracked (block **1579**).

[0364] Training

[0365] As discussed above, each machine learning model may be trained on specific aspects of the normal pattern of life for the system such as devices, users, network traffic flow, outputs from one or more cyber security analysis tools analyzing the system, email contact associations for each user, email characteristics, etc. The one or more machine learning models may use at least unsupervised learning algorithms to establish what is the normal pattern of life for the system. The machine learning models can train on both (i) the historical normal distribution of alerts and events for that system as well as (ii) factored in is a normal distribution information from similar peer systems to establish the normal pattern of life of the behavior of alerts and/or events for that system. Another set of machine learning models can train on characteristics of emails and the activities and behavior of its email users to establish a normal for these.

[0366] The models may leverage at least two different approaches to detecting anomalies (e.g., comparing each system's behavior to its own history, comparing that system to its peers' history, and/or comparing an email to both characteristics of emails and the activities and behavior of its email users). This multiple source comparison may allow the models to avoid learning existing bad behavior as 'a normal' because compromised devices/users/components/emails will exhibit behavior different to their immediate peers.

[0367] In addition, the one or more machine learning models can use the comparison of (i) the normal pattern of life for that system corresponding to the historical normal distribution of alerts and events for that system mapped out

in the same multiple dimension space to (ii) the current chain of individual alerts and events behavior under analysis. This comparison can yield detection of the one or more unusual patterns of behavior within the plotted individual alerts and/or events, which allows the detection of previously unidentified cyber threats compared to finding cyber threats with merely predefined descriptive objects and/or signatures. Thus, increasingly intelligent malicious cyber threats that try to pick and choose when they take their actions in order to generate low level alerts and events will still be detected, even though they have not yet been identified by other methods of cyber analysis. These intelligent malicious cyber threats can include malware, spyware, key loggers, malicious links in an email, malicious attachments in an email, etc. as well as nefarious internal information technology staff who know intimately how to not set off any high level alerts or events.

[0368] In essence, the plotting and comparison may be a way to filter out what is normal for that system and then may be able to focus the analysis on what is abnormal or unusual for that system. Then, for each hypothesis of what could be happening with the chain of unusual events and/or alerts, the gatherer module may gather additional metrics from the data store including the pool of metrics originally considered 'normal behavior' to support or refute each possible hypothesis of what could be happening with this chain of unusual behavior under analysis.

[0369] Each of the individual alerts and/or events in a chain of alerts and/or events that form the unusual pattern can indicate subtle abnormal behavior; and thus, each alert and/or event can have a low threat risk associated with that individual alert and/or event. However, when analyzed as a distinct chain/grouping of alerts and/or events behavior forming the chain of unusual pattern by the one or more machine learning models, then that distinct chain of alerts and/or events can be used to determine to now have a much higher threat risk than any of the individual alerts and/or events in the chain.

[0370] In addition, new cyberattacks can be of such severity and speed that a human response cannot happen quickly enough. Thanks to these self-learning advances, it is now possible for a machine to uncover these emerging threats and deploy appropriate, real-time responses to fight back against the most serious cyber threats.

[0371] The threat detection system may have the ability to self-learn and detect normality in order to spot true anomalies, allowing organizations of all sizes to understand the behavior of users and machines on their networks at both an individual and group level. Monitoring behaviors, rather than using predefined descriptive objects and/or signatures, may mean that more attacks can be spotted ahead of time and extremely subtle indicators of wrongdoing can be detected. Unlike traditional legacy defenses, a specific attack type or new malware does not have to have been seen first before it can be detected. A behavioral defense approach mathematically models both machine, email, and human activity behaviorally, at and after the point of compromise, in order to predict and catch today's increasingly sophisticated cyber-attack vectors. It is thus possible to computationally establish what is normal, in order to then detect what is abnormal. In addition, the machine learning constantly revisits assumptions about behavior, using probabilistic mathematics. The cyber-threat defense system's unsupervised machine learning methods do not require training data with pre-defined labels. Instead, they are able to identify key patterns and trends in the data, without the need for human input.

[0372] The cyber-threat defense system **100** may use at least three separate machine-learning models. Each machine-learning model may be trained on specific aspects of the normal pattern of life for the system such as devices, users, network traffic flow, outputs from one or more cyber security analysis tools analyzing the system, etc. One or more machine learning models may also be trained on characteristics and aspects of all manner of types of cyber threats. One or more machine learning models may also be trained on characteristics of emails themselves. In an embodiment, the one or more models may be trained on specific aspects of these broader concepts. For example, the models may be specifically trained on associations, attachments, compliances, data loss & transfers, general, metadata, hygiene, links, proximity, spoof, type, validation, and other anomalies.

[0373] In addition, the one or more machine learning models can be self-learning using unsupervised learning algorithms. For example, a set of the one or more machine learning models can be trained on the normal behavior of users and their emails use data from the probes to train on; and therefore, regularly update what a base line for the normal behavior is. This autonomous, self-learning defense system protects against malicious activity in the email domain—whether the malicious activity is from any of (i) standard threat actors from email, such as phishing and malware emails, and (ii) insider threat from users, which does not rely solely on pre-recognized, arbitrary ideas of malicious email domain activity but instead autonomously contextualizes each communication to assess its anomaly compared to standard behavior of the user and organization.

Clauses

[0374] Embodiments of the above disclosure can be described with reference to the following numbered clauses, with preferred features laid out in the dependent clauses:

[0375] Clause 1. An apparatus for determining and acting on a cyber threat risk of a structured document addressed to a recipient by a sender, the apparatus comprising one or more machine learning models that are trained on the classification of structured documents with one or more of a plurality of categories based on a plurality of characteristics of the structured documents; a classifier configured to receive a structured document for analysis and to parse the structured document to extract the plurality of characteristics of the structured document; wherein the classifier is further configured to classify the structured document with one or more of the plurality of categories based on the extracted plurality of characteristics and the one or more machine learning models, and to determine an associated score for the classification; and an autonomous response module configured to, based on a comparison of the associated score with a threshold, cause one or more autonomous actions to be taken in relation to the structured document.

[0376] Clause 2. The apparatus of clause 1, further comprising a sender user interface; wherein the structured document is to be sent from the sender to an indicated recipient; wherein each category of the plurality of categories represents a respective recipient of a plurality of recipients known to the sender; wherein the associated score represents the probability of a match between the indicated recipient and

the extracted plurality of characteristics; wherein the classifier is further configured to determine one or more further scores representing the respective probability of a match between the extracted plurality of characteristics and each of the other recipients known to the sender; wherein the threshold represents the score of an alternative recipient, of the other recipients known to the sender, having the highest probability of a match; and wherein the one or more autonomous actions comprise, if the associated score is less than the threshold, displaying an alert to the sender on the sender user interface indicating that the alternative recipient has a higher probability of a match than the indicated recipient.

[0377] Clause 3. The apparatus of clause 2, wherein one or more of the machine learning models have been trained to identify, for each recipient known to the sender, one or more indicators corresponding to characteristics that are frequently present in structured documents sent by the sender and addressed to the respective recipient known to the sender relative to those addressed to other recipients known to the sender; and wherein the classifier is configured to classify the structured document with one or more of the categories representing the plurality of recipients known to the sender by comparing the extracted plurality of characteristics with the one or more indicators for each recipient known to the sender.

[0378] Clause 4. The apparatus of clause 3, wherein the training of the one or more machine learning models is configured to be periodically updated by further training with the classification of structured documents sent by the sender since the last update.

[0379] Clause 5. The apparatus of any of clauses 2 to 4, wherein the autonomous response module is only configured to cause the alert to be displayed if the associated score is less than the threshold by more than a given amount.

[0380] Clause 6. The apparatus of any of clauses 2 to 5, wherein the autonomous response module is further configured to prevent the structured document from being sent to the indicated recipient until the alert has been acknowledged by the sender.

[0381] Clause 7. The apparatus of any of clauses 3 to 6, wherein the one or more machine learning modules have been trained to filter out indicators that are identified as common based on a structured document training data set that includes structured documents from a plurality of senders.

[0382] Clause 8. The apparatus of any of clauses 3 to 7, wherein the plurality of characteristics of the structured document comprises the constituent words and/or phrases of a body text of the structured document.

[0383] Clause 9. The apparatus of any of clauses 3 to 7, wherein the plurality of characteristics of the structured document comprises the stem of the constituent words and/or phrases of a body text of the structured document.

[0384] Clause 10. The apparatus of any of clauses 3 to 9, wherein the plurality of characteristics of the structured document further comprise additional recipients indicated in the structured document to be sent from the sender; and wherein the one or more indicators for each respective recipient known to the sender further comprise additional recipients that are frequently present in structured documents sent by the sender and addressed to the respective recipient relative to those addressed to other recipients known to the sender.

[0385] Clause 11. The apparatus of any of clauses 3 to 10, wherein the sender is associated with an organization, and wherein the classifier is further configured to classify the structured document with the one or more of the categories representing the plurality of recipients known to the sender by comparing the extracted plurality of characteristics with one or more additional indicators, the additional indicators corresponding to characteristics that are frequently present in structured documents sent by other senders associated with the organization and addressed to the respective recipients; wherein the one or more additional indicators are weighted lower than the one or more indicators in the classification.

[0386] Clause 12. The apparatus of any of clauses 3 to 10, wherein the sender is associated with an organization; wherein the classifier is further configured to classify the structured document with one or more categories representing unknown recipients that are unknown to the sender based on unique indicators corresponding to characteristics that are uniquely present in structured documents sent to the respective unknown recipients by other senders associated with the organization; and wherein the one or more autonomous actions comprise, if the score associated with the unknown recipient corresponds to the highest probability of a match, displaying an alert to the sender on the sender user interface indicating that the unknown recipient has a higher probability of a match than the indicated recipient.

[0387] Clause 13. The apparatus of any of clauses 3 to 12, wherein the autonomous response module is further configured to display, with the alert to the sender, one or more of the characteristics and/or indicators that led to the alternative recipient having a higher probability of a match.

[0388] Clause 14. The apparatus of clause 1, wherein the structured document has been sent to the user from a given sender; wherein the one or more categories comprise one or more malign categories; and wherein, when the associated score determined for the one or more malign categories is above the threshold, the one or more autonomous actions comprise one or more actions to contain the malign nature of the sent structured document.

[0389] Clause 15. The apparatus of clause 14, wherein the plurality of characteristics of the structured document comprise one or more of: the constituent words and/or phrases of a body text of the structured document, links in the structured document directing to other resources, attachments of the structured document, a format of an addressing field of the structured document; the presence of phone numbers in the body text, the presence of email addresses in the body text, the presence of currency values in the body text, and/or derived ratio analysis of aspects of text construction of the body text.

[0390] Clause 16. The apparatus of clause 15, further comprising a language classifier with one or more language machine learning models trained to identify a language of text; wherein the one or more of the machine learning models trained on the classification of structured documents are trained on words and/or phrases of a subset of languages; wherein the language classifier is configured to reference the one or more language machine learning modules to identify the language of the body text of the structured document; and wherein, if the language of the body text is determined to be a language not included in the subset of languages, the classifier is configured to classify the structured document with one or more of the plurality of categories based on the

extracted plurality of characteristics excluding the constituent words and/or phrases of a body text of the structured document.

[0391] Clause 17. The apparatus of any of clauses 14 to 16, wherein the one or more actions to contain the malign nature of the received structured document comprise one or more of: preventing the delivery of the structured document to the user, removing the structured document from a user inbox; converting one or more attachments of the structured document from one file format to another file format, removing one or more attachments of the structured document, redirecting links in the structured document to alternative destinations, removing links from the structured document, tagging the structured document as junk, redirecting or copying the structured document to another user inbox, inserting additional text into the structured document, and/or altering the content of one or more defined fields of the structured document.

[0392] Clause 18. The apparatus of any of clauses 14 to 17, further comprising a user interface having an administrative tool for setting, by a user, which types of autonomous actions the autonomous response module is configured to perform and for setting the threshold.

[0393] Clause 19. The apparatus of any of clauses 14 to 18, wherein the one or more machine learning models that are trained on the classification of structured documents comprise at least one machine learning module trained by comparing the relative frequency density of words and phrases in training data sets corresponding to each respective category.

[0394] Clause 20. The apparatus of any of clauses 14 to 19, further comprising a user interface wherein the autonomous response module is further configured to display, on the user interface, one or more of the characteristics of the structured document that led to the cause of the autonomous action.

[0395] Clause 21. A computer implemented method for determining and acting on a cyber threat risk of a structured document addressed to a recipient by a sender, the method comprising: using one or more machine learning models that are trained on the classification of structured documents with one or more of a plurality of categories based on a plurality of characteristics of the structured documents; receiving, at a classifier, a structured document for analysis and parsing the structured document to extract the plurality of characteristics of the structured document; classifying, at the classifier, the structured document with one or more of the plurality of categories based on the extracted plurality of characteristics and the one or more machine learning models, and determining an associated score for the classification; and causing, by an autonomous response module, one or more autonomous actions to be taken in relation to the structured document based on a comparison of the associated score with a threshold.

[0396] Clause 22. The computer implemented method of clause 21, wherein the structured document is to be sent from the sender to an indicated recipient; wherein each category of the plurality of categories represents a respective recipient of a plurality of recipients known to the sender; and wherein the associated score represents the probability of a match between the indicated recipient and the extracted plurality of characteristics; the computer implemented method further comprising: determining, by the classifier, one or more further scores representing the respective probability of a match between the extracted plurality of characteristics and each of the other recipients known to the sender; wherein the threshold represents the score of an alternative recipient, of the other recipients known to the sender, having the highest probability of a match; and wherein the one or more autonomous actions comprise, if the associated score is less than the threshold, displaying an alert to the sender on the sender user interface indicating that the alternative recipient has a higher probability of a match than the indicated recipient.

[0397] Clause 23. The computer implemented method of clause 22, wherein one or more of the machine learning models have been trained to identify, for each recipient known to the sender, one or more indicators corresponding to characteristics that are frequently present in structured documents sent by the sender and addressed to the respective recipient known to the sender relative to those addressed to other recipients known to the sender; and wherein the classifier classifies the structured document with one or more of the categories representing the plurality of recipients known to the sender by comparing the extracted plurality of characteristics with the one or more indicators for each recipient known to the sender.

[0398] Clause 24. The computer implemented method of clause 23, wherein the training of the one or more machine learning models is periodically updated by further training with the classification of structured documents sent by the sender since the last update.

[0399] Clause 25. The computer implemented method of any of clauses 22 to 24, wherein the autonomous response module is only configured to cause the alert to be displayed if the associated score is less than the threshold by more than a given amount.

[0400] Clause 26. The computer implemented method of any of clauses 22 to 25, wherein the autonomous response module prevents the structured document from being sent to the indicated recipient until the alert has been acknowledged by the sender.

[0401] Clause 27. The computer implemented method of any of clauses 23 to 26, wherein the one or more machine learning modules have been trained to filter out indicators that are identified as common based on a structured document training data set that includes structured documents from a plurality of senders.

[0402] Clause 28. The computer implemented method of any of clauses 23 to 27, wherein the plurality of characteristics of the structured document comprises the constituent words and/or phrases of a body text of the structured document.

[0403] Clause 29. The computer implemented method of any of clauses 23 to 27, wherein the plurality of characteristics of the structured document comprises the stem of the constituent words and/or phrases of a body text of the structured document.

[0404] Clause 30. The computer implemented method of any of clauses 23 to 29, wherein the plurality of characteristics of the structured document further comprise additional recipients indicated in the structured document to be sent from the sender; and wherein the one or more indicators for each respective recipient known to the sender further comprise additional recipients that are frequently present in structured documents sent by the sender and addressed to the respective recipient relative to those addressed to other recipients known to the sender.

[0405] Clause 31. The computer implemented method of any of clauses 23 to 30, wherein the sender is associated with an organization, and wherein the classifier classifies the structured document with the one or more of the categories representing the plurality of recipients known to the sender by comparing the extracted plurality of characteristics with one or more additional indicators, the additional indicators corresponding to characteristics that are frequently present in structured documents sent by other senders associated with the organization and addressed to the respective recipients; wherein the one or more additional indicators are weighted lower than the one or more indicators in the classification.

[0406] Clause 32. The computer implemented method of any of clauses 23 to 30, wherein the sender is associated with an organization; wherein the classifier classifies the structured document with one or more categories representing unknown recipients that are unknown to the sender based on unique indicators corresponding to characteristics that are uniquely present in structured documents sent to the respective unknown recipients by other senders associated with the organization; and wherein the one or more autonomous actions comprise, if the score associated with the unknown recipient corresponds to the highest probability of a match, displaying an alert to the sender on the sender user interface indicating that the unknown recipient has a higher probability of a match than the indicated recipient.

[0407] Clause 33. The computer implemented method of any of clauses 23 to 32, further comprising displaying with the alert to the sender, by the autonomous response module, one or more of the characteristics and/or indicators that led to the alternative recipient having a higher probability of a match.

[0408] Clause 34. The computer implemented method of clause 21, wherein the structured document has been sent to the user from a given sender; wherein the one or more categories comprise one or more malign categories; and wherein, when the associated score determined for the one or more malign categories is above the threshold, the one or more autonomous actions comprise one or more actions to contain the malign nature of the sent structured document.

[0409] Clause 35. The computer implemented method of clause 24, wherein the plurality of characteristics of the structured document comprise one or more of: the constituent words and/or phrases of a body text of the structured document, links in the structured document directing to other resources, attachments of the structured document, a format of an addressing field of the structured document; the presence of phone numbers in the body text, the presence of email addresses in the body text, the presence of currency values in the body text, and/or derived ratio analysis of aspects of text construction of the body text.

[0410] Clause 36. The computer implemented method of clause 35, further comprising a language classifier with one or more language machine learning models trained to identify a language of text; wherein the one or more of the machine learning models trained on the classification of structured documents are trained on words and/or phrases of a subset of languages; the computer implemented method further comprising: referencing, by the language classifier, the one or more language machine learning modules to identify the language of the body text of the structured document; and wherein, if the language of the body text is determined to be a language not included in the subset of languages, classifying the structured document with one or more of the plurality of categories based on the extracted plurality of characteristics excluding the constituent words and/or phrases of a body text of the structured document.

[0411] Clause 37. The computer implemented method of any of clauses 34 to 36, wherein the one or more actions to contain the malign nature of the received structured document comprise one or more of: preventing the delivery of the structured document to the user, removing the structured document from a user inbox; converting one or more attachments of the structured document from one file format to another file format, removing one or more attachments of the structured document, redirecting links in the structured document to alternative destinations, removing links from the structured document, tagging the structured document as junk, redirecting or copying the structured document to another user inbox, inserting additional text into the structured document, and/or altering the content of one or more defined fields of the structured document.

[0412] Clause 38. The computer implemented method of any of clauses 34 to 37, wherein the threshold and the types of autonomous actions the autonomous response module perform are set by a user in a user interface having an administrative tool.

[0413] Clause 39. The computer implemented method of any of clauses 34 to 38, wherein the one or more machine learning models that are trained on the classification of structured documents comprise at least one machine learning module trained by comparing the relative frequency density of words and phrases in training data sets corresponding to each respective category.

[0414] Clause 40. The computer implemented method of any of clauses 34 to 39, further comprising displaying one or more of the characteristics of the structured document that led to the cause of the autonomous action on a user interface.

[0415] Clause 41. A non-transitory computer readable medium including executable instructions that, when executed with one or more processors, cause a cyber defense system to perform the operations of clause 21.

[0416] In one aspect of the disclosure, an apparatus for determining and acting on a cyber threat risk of a structured document to be sent from a sender to an indicated recipient is provided, the apparatus comprising: one or more machine learning models that are trained on the classification of structured documents, with one or more of a plurality of recipients known to the sender, based on a plurality of characteristics of the structured documents; a classifier configured to receive a structured document for analysis and to parse the structured document to extract the plurality of characteristics of the structured document; and to use the one or more machine learning models to classify the structured document by determining a set of respective match probability scores between the extracted plurality of characteristics and each of the known recipients, including the indicated recipient; a sender user interface; and an autonomous response module configured to, determine an expected recipient corresponding to the known recipient having the highest probability of a match; and, if the indicated recipient is not the expected recipient, to cause an alert to be displayed to the sender on the sender user interface indicating that the expected recipient has a higher probability of a match with the structured document than the indicated recipient.

[0417] While the foregoing design and embodiments thereof have been provided in considerable detail, it is not

the intention of the applicant(s) for the design and embodiments provided herein to be limiting. Additional adaptations and/or modifications are possible, and, in broader aspects, these adaptations and/or modifications are also encompassed. Accordingly, departures may be made from the foregoing design and embodiments without departing from the scope afforded by the following claims, which scope is only limited by the claims when appropriately construed.

What is claimed is:

1. A cyber security appliance configurable to be coupled to a computer system, comprising:

an email module configured to (i) process incoming emails and (ii) log data and metadata associated with the incoming emails;

a cyber threat module coupled to the email module and configured to analyze the logged data and metadata to assess a severity level of a cyber threat using one or more Artificial Intelligence (AI) models trained on one or more of the group of parameters consisting of: (i) email metrics, (ii) suspicious characteristics of emails, (iii) malicious links and attachments, (iv) a database of known malicious actors, and (v) any combination of these parameters, in order to cluster emails with similar parameters;

an AI classifier module coupled to the email module and the cyber threat module and configured to determine the likelihood of an email cyber threat campaign is occurring using one or more AI models and analysis of the parameters;

an autonomous response module coupled to the email module, the cyber threat module, and the AI classifier module and configured to act against an email determined to be a threat when instructed by at least one of the cyber threat module and the AI classifier module, where the autonomous response module is configured to take an autonomous action against the email, where the email acted against had previously been examined and not stopped or acted upon but now is deemed part of an email campaign; and

a user interface module coupled to the email module, the cyber threat module, the AI classifier module, and the autonomous response module and configured to perform at least one of the actions in the group consisting of: (i) generating a report, (ii) presenting data on a display, and (iii) showing a graphical display of the system indicating the details of a cyber threat campaign.

2. The cyber security appliance of claim **1**, wherein the email module is further configured to detect incoming email cyber threats using one or more Artificial Intelligence (AI) models trained on the normal pattern of life in the computer system.

3. The cyber security appliance of claim **1**, wherein the suspicious characteristics of emails consists of at least one of the group consisting of: (i) the email wording, (ii) the geographic location of a host or a domain, (iii) the relationship between sender and recipient, and (iv) the presence of a malign inducement.

4. The cyber security appliance of claim **1**, wherein the AI classifier module is further configured to perform one or more of the group consisting of: (i) scoring emails according to a plurality of indices, (ii) grouping similar emails into clusters, (iii) examining the chronological sequence of detected cyber threats for fluctuations, (iv) analyzing each mailbox to determine its purpose, and (v) estimating the probability of unsolicited emails to each mailbox, and where the autonomous action is taken retrospectively on that email in order to have a positive impact of potentially stopping any harm from the email campaign.

5. The cyber security appliance of claim **4**, wherein the AI classifier module is further configured to find outliers based on the detected cyber threats fluctuations and to return the outliers to one or more AI models for training, and where the user interface module is further configured to display on a display screen trends for one or more ongoing email campaign detections, and the emails involved in the email campaign and the email accounts that have been targeted.

6. The cyber security appliance of claim **4**, wherein the AI classifier module is further configured to perform at least one of the group consisting of (i) determining if a mailbox is public-facing or private-facing, (ii) logging the number of detected threats, and (iii) continuously scoring each mailbox based on the occurrence of threats to that mailbox.

7. The cyber security appliance of claim **1**, wherein the autonomous response module is configured to take the autonomous action against detected cyber threat emails and email cyber campaigns by one or more of the actions in the group consisting of: (i) converting an attachment into a harmless format, (ii) stripping an attachment from an email, (iii) unspoofing an email header, (iv) deleting a link, (v) locking a link, (vi) double locking a link, (vii) holding a message from an email inbox, and (viii) moving an email to a junk email box.

8. The cyber security appliance of claim **1**, wherein the autonomous response module is configured to take the autonomous action against detected cyber threat emails and email cyber campaigns is one or more of the group consisting of: (i) converting an attachment into a harmless format, (ii) stripping an attachment from an email, (iii) unspoofing an email header, (iv) deleting a link, (v) locking a link, (vi) double locking a link, (vii) holding a message from an email inbox, and (viii) moving an email to a junk email box.

9. The cyber security appliance of claim **1**, wherein the user interface module is further configured to perform at least one of the group consisting of: (i) providing a single plane of analysis, (ii) displaying the risk profile of entities, (iii) tracking high permission users, (iv) tracking SaaS documents, (v) tracking macro-enabled documents, (vi) tracking documents within an organization, and (vii) model a connectivity web.

10. The cyber security appliance of claim **1**, wherein the cyber security appliance is further configured to coordinate actions and share information with a second cyber security appliance coupled to the computer system in order to communicate when an email campaign is detected.

11. A method of operating a cyber security appliance configurable to be coupled to a computer system, comprising:

processing incoming emails;

logging data and metadata associated with the incoming emails;

determining a level of email cyber threat using one or more AI modules trained on one or more of the group of parameters consisting: of (i) email metrics, (ii) suspicious characteristics of emails, (iii) malicious links and attachments, (iv) a database of known malicious actors to assess a severity level of a cyber threat,

and (v) any combination of these parameters, in order to cluster emails with similar parameters;

assessing the likelihood of an email cyber threat campaign is occurring using one or more AI models and analysis of the parameters;

taking autonomous action against an email determined to be a threat when instructed by at least one of the cyber threat module and the AI classifier module, where the autonomous response module is configured to take an autonomous action against the email, where the email acted against had previously been examined and not stopped or acted upon but now is deemed part of an email campaign; and

performing at least one of the reporting actions in the group consisting of (i) generating a report, (ii) presenting data on a screen, and (iii) showing a graphical display of the system indicating the details of a cyber threat campaign.

12. The method of claim **11**, wherein the processing of all incoming emails uses one or more Artificial Intelligence (AI) models trained on the normal pattern of life in the computer system to detect incoming email cyber threats.

13. The method of claim **11**, wherein the suspicious characteristics of emails consists of at least one of the group consisting of: (i) the email wording, (ii) the geographic location of a host or a domain, (iii) the relationship between sender and recipient, and (iv) the presence of a malign inducement.

14. The method of claim **11**, wherein the assessing the likelihood of an email cyber threat campaign further comprises one or more of the group consisting of: (i) scoring emails according to a plurality of indices, (ii) grouping similar emails into clusters, (iii) examining the chronological sequence of detected cyber threats for fluctuations, (iv) analyzing each mailbox to determine its purpose, and (v) estimating the probability of unsolicited emails to each mailbox, and where the autonomous action is taken retrospectively on that email in order to have a positive impact of potentially stopping any harm from the email campaign.

15. The method of claim **14**, wherein the cyber security appliance is further configured to find outliers based on the detected cyber threats fluctuations and to return the outliers to one or more AI models for training, and where the user interface module is further configured to display on a display screen trends for one or more ongoing email campaign detections, and the emails involved in the email campaign and the email accounts that have been targeted.

16. The method of claim **14**, wherein the cyber security appliance is further configured to perform at least one of the group consisting of (i) determining if a mailbox is public-facing or private-facing, (ii) logging the number of detected threats, and (iii) continuously scoring each mailbox based on the occurrence of threats to that mailbox.

17. The method of claim **11**, wherein assessing the likelihood that an email cyber threat campaign is occurring further comprises (i) comparing a detected email cyber campaign to previously detected cyber email campaigns and (ii) analyzing a detected email cyber campaign to understand its origin and purpose.

18. The method of claim **11**, wherein taking the autonomous action against detected email cyber threats and email cyber campaigns further comprises one or more of the group of actions consisting of: (i) converting an attachment into a harmless format, (ii) stripping an attachment from an email, (iii) unspoofing an email header, (iv) deleting a link, (v) locking a link, (vi) double locking a link, (vii) holding a message from an email inbox, and (viii) moving an email to a junk email box.

19. The method of claim **11**, wherein performing at least one of the reporting actions further comprises (i) providing a single plane of analysis, (ii) displaying the risk profile of entities, (iii) tracking high permission users, (iv) tracking SaaS documents, (v) tracking macro-enabled documents, (vi) tracking documents within an organization, and (vii) model a connectivity web.

20. A non-transitory computer-readable medium comprising computer-readable code operable, when executed by one or more processing apparatuses in a computer system, to instruct a computing device to perform the method of claim **11**.

* * * * *