



(19) **United States**  
(12) **Patent Application Publication**  
LAVI et al.

(10) **Pub. No.: US 2023/0171278 A1**  
(43) **Pub. Date: Jun. 1, 2023**

(54) **DARK WEB MONITORING, ANALYSIS AND ALERT SYSTEM AND METHOD**

*G06F 16/2457* (2006.01)  
*G06F 16/951* (2006.01)

(71) Applicant: **Sixgill Ltd.**, Tel Aviv (IL)  
(72) Inventors: **Elad LAVI**, Ramat HaKovesh (IL); **Avi KASZTAN**, Petah Tikva (IL)

(52) **U.S. Cl.**  
CPC ..... *H04L 63/1425* (2013.01); *G06F 16/285* (2019.01); *G06F 16/951* (2019.01); *G06F 16/24578* (2019.01); *G06F 21/00* (2013.01); *G06Q 10/063* (2013.01); *G06Q 50/26* (2013.01); *H04L 63/0227* (2013.01); *H04L 63/1408* (2013.01); *H04W 4/21* (2018.02)

(73) Assignee: **Sixgill Ltd.**, Tel Aviv (IL)

(21) Appl. No.: **18/098,770**

(22) Filed: **Jan. 19, 2023**

(57) **ABSTRACT**

**Related U.S. Application Data**

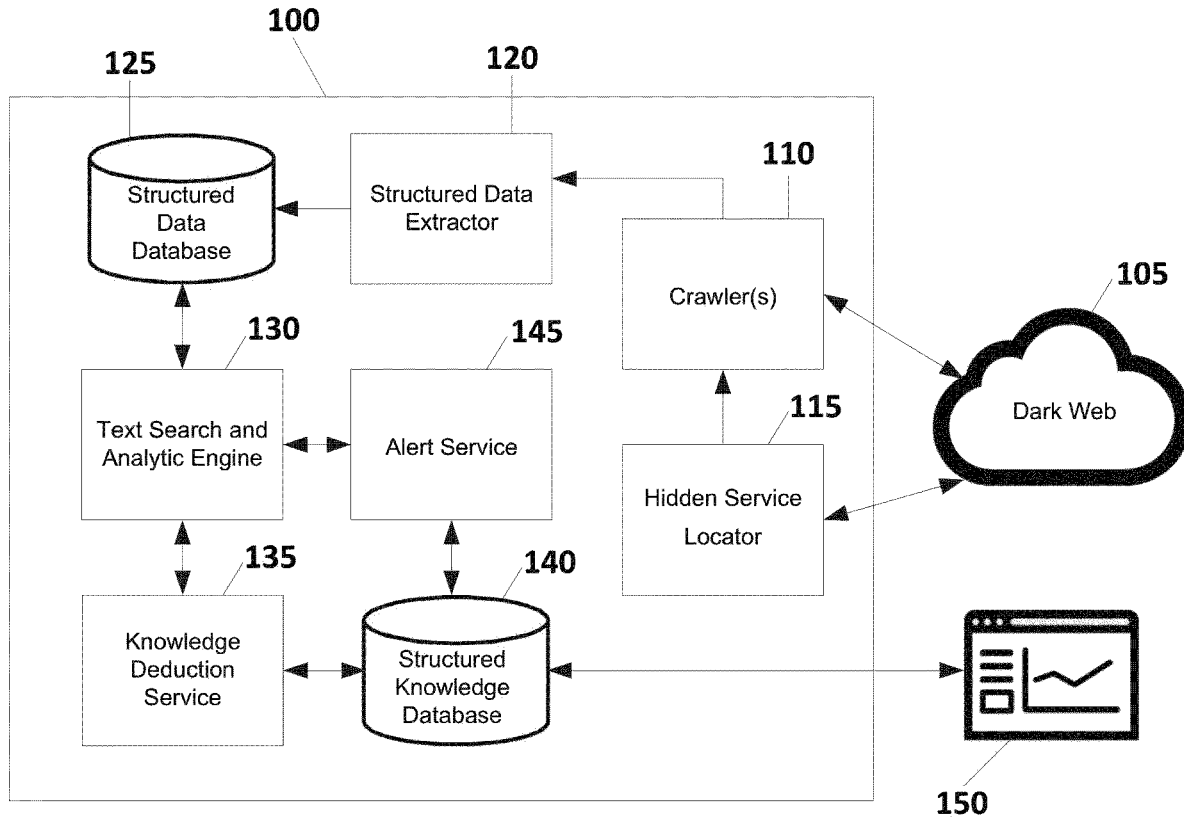
(63) Continuation of application No. 16/066,315, filed as application No. PCT/IB2016/058016 on Dec. 27, 2016, now Pat. No. 11,570,188.

A dark web monitoring, analysis and alert system comprising a data receiving module configured to receive data collected from the dark web and structured; a Structured Data Database (SDD) connected with the data receiving module, the SDD configured to store the structured data; a Text Search and Analytic Engine (TSAE) connected with the SDD, the TSAE configured to enable advanced search and basic analysis in the structured data; a Knowledge Deduction Service (KDS) connected with the TSAE, the KDS configured to deeply analyze the collected data; the deep analysis comprises extracting insights regarding dark web surfers behavioral patterns and interactions; a Structured Knowledge Database (SKD) connected with the KDS, the SKD configured to store the deep analysis results; and an Alert Service connected with the TSAE and the SKD, the Alert Service configured to provide prioritized alerts based on the deep analysis.

(60) Provisional application No. 62/271,344, filed on Dec. 28, 2015.

**Publication Classification**

(51) **Int. Cl.**  
*H04L 9/40* (2006.01)  
*G06Q 50/26* (2006.01)  
*G06F 21/00* (2006.01)  
*H04W 4/21* (2006.01)  
*G06Q 10/063* (2006.01)  
*G06F 16/28* (2006.01)



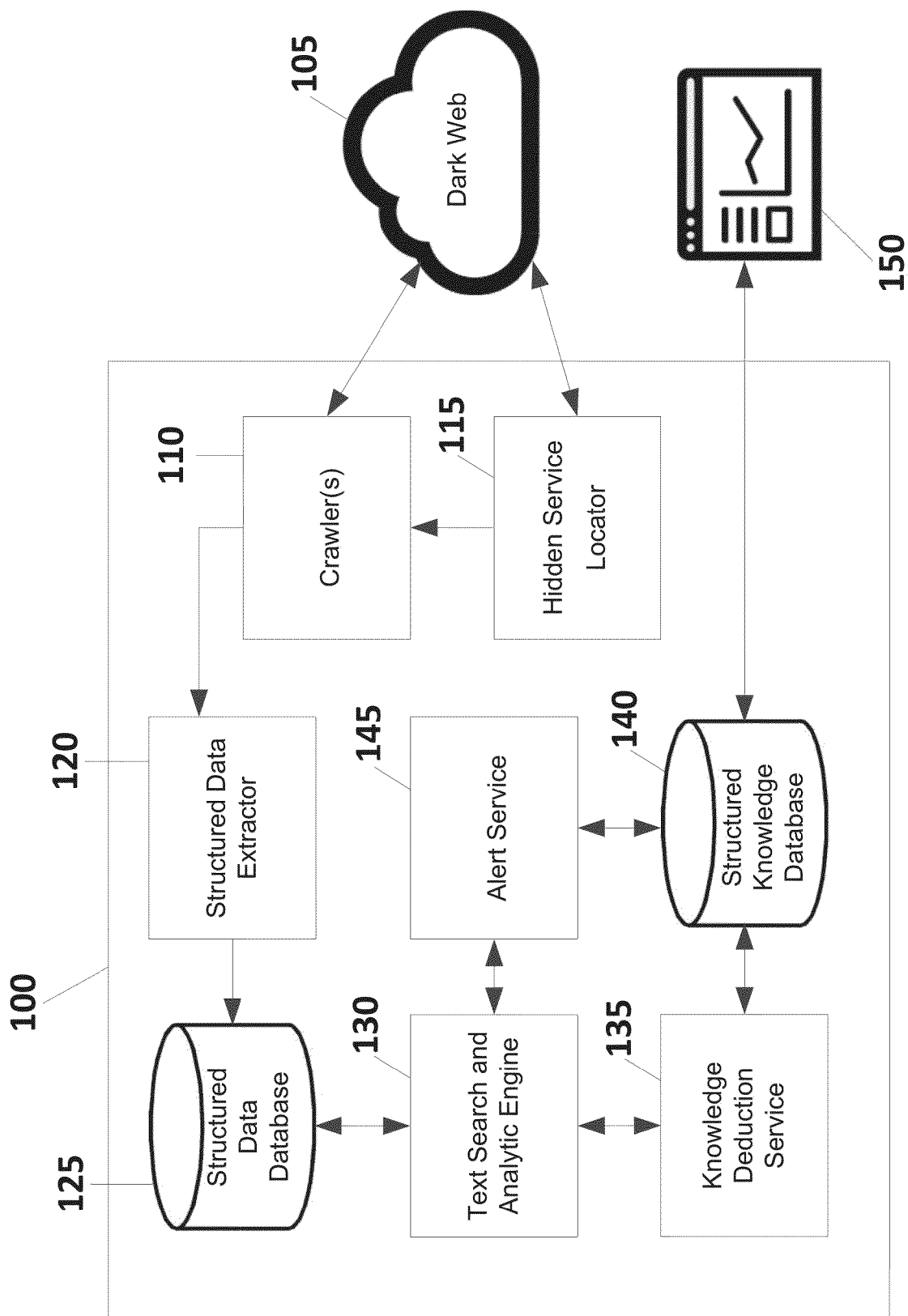


Fig. 1

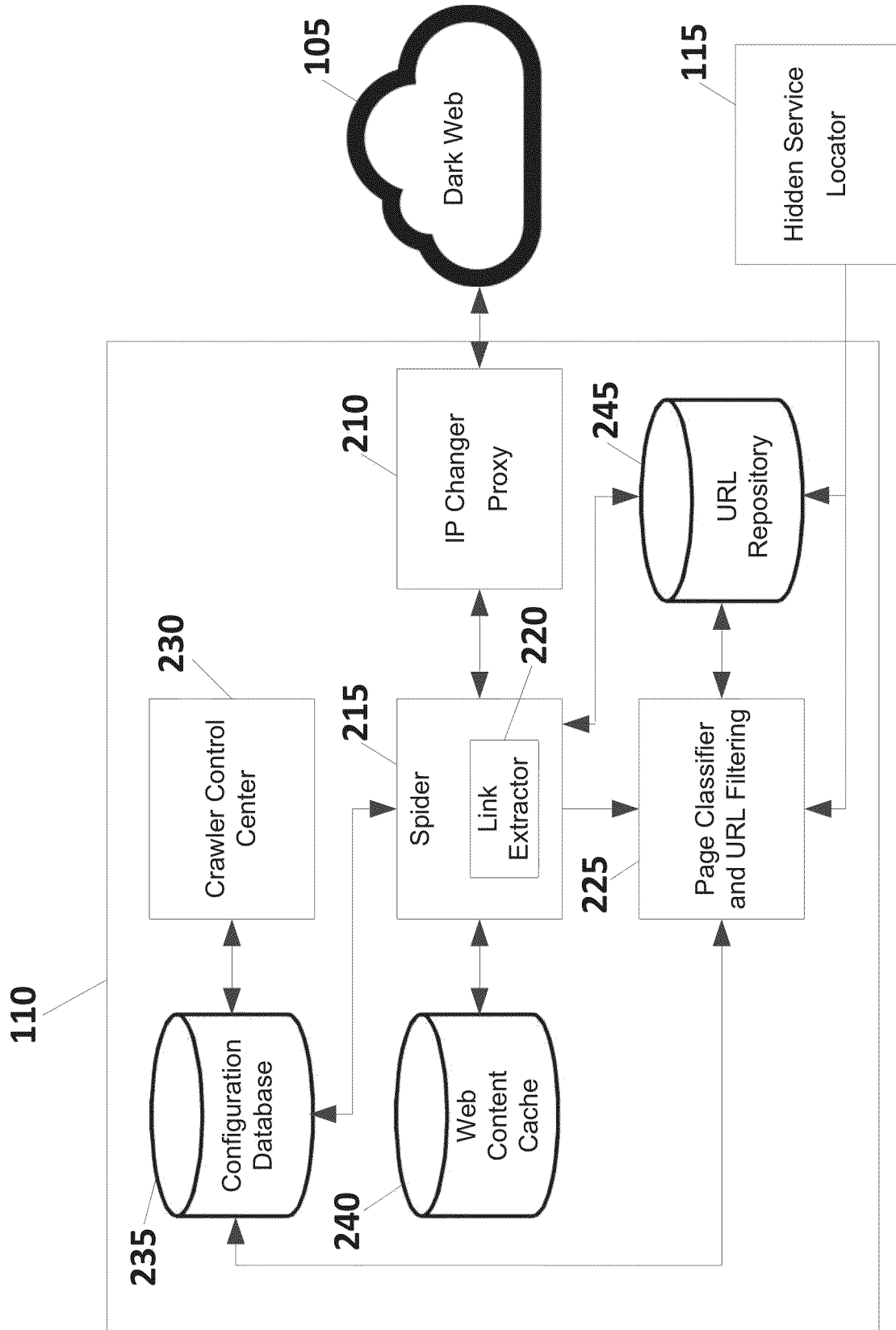
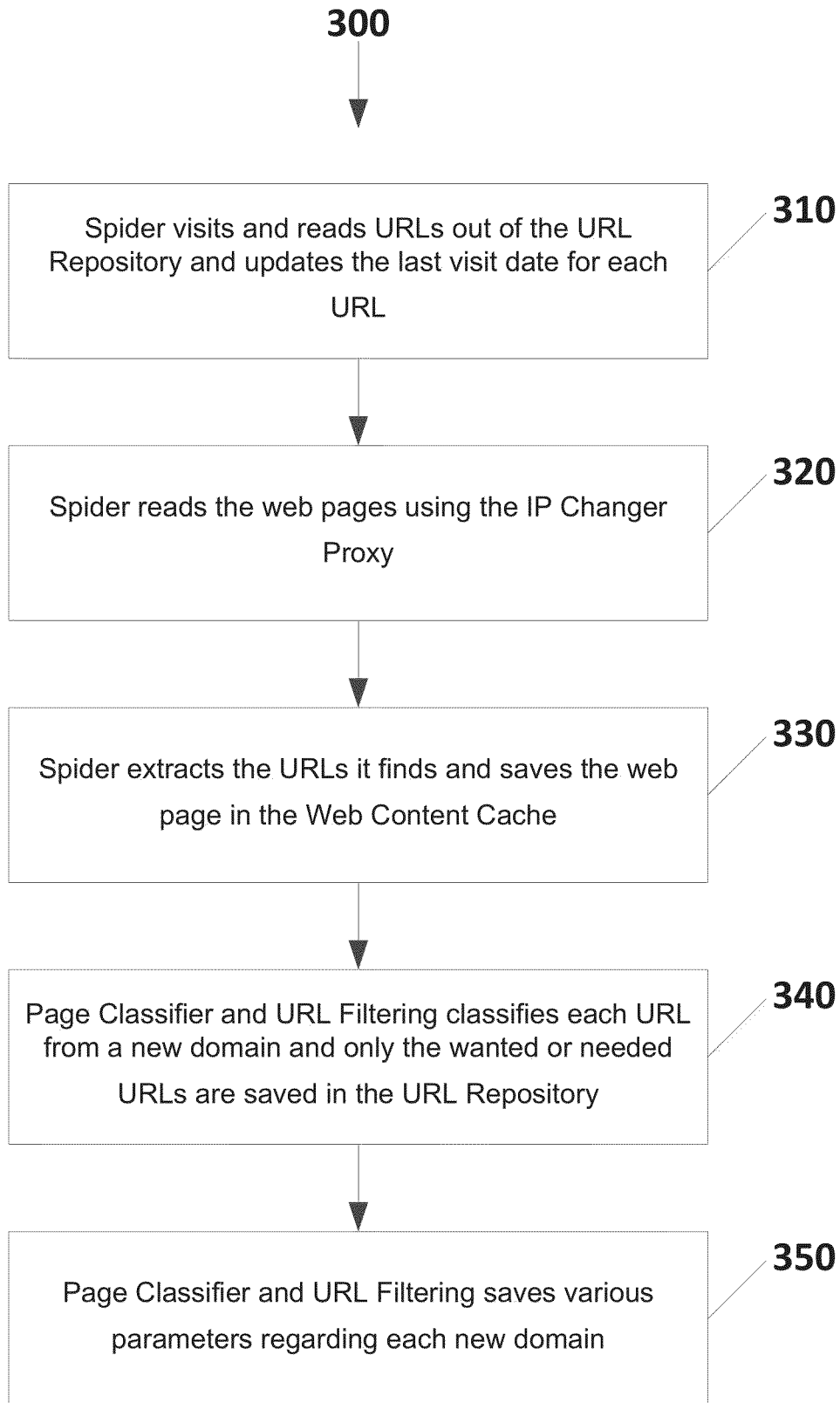


Fig. 2



**Fig. 3**

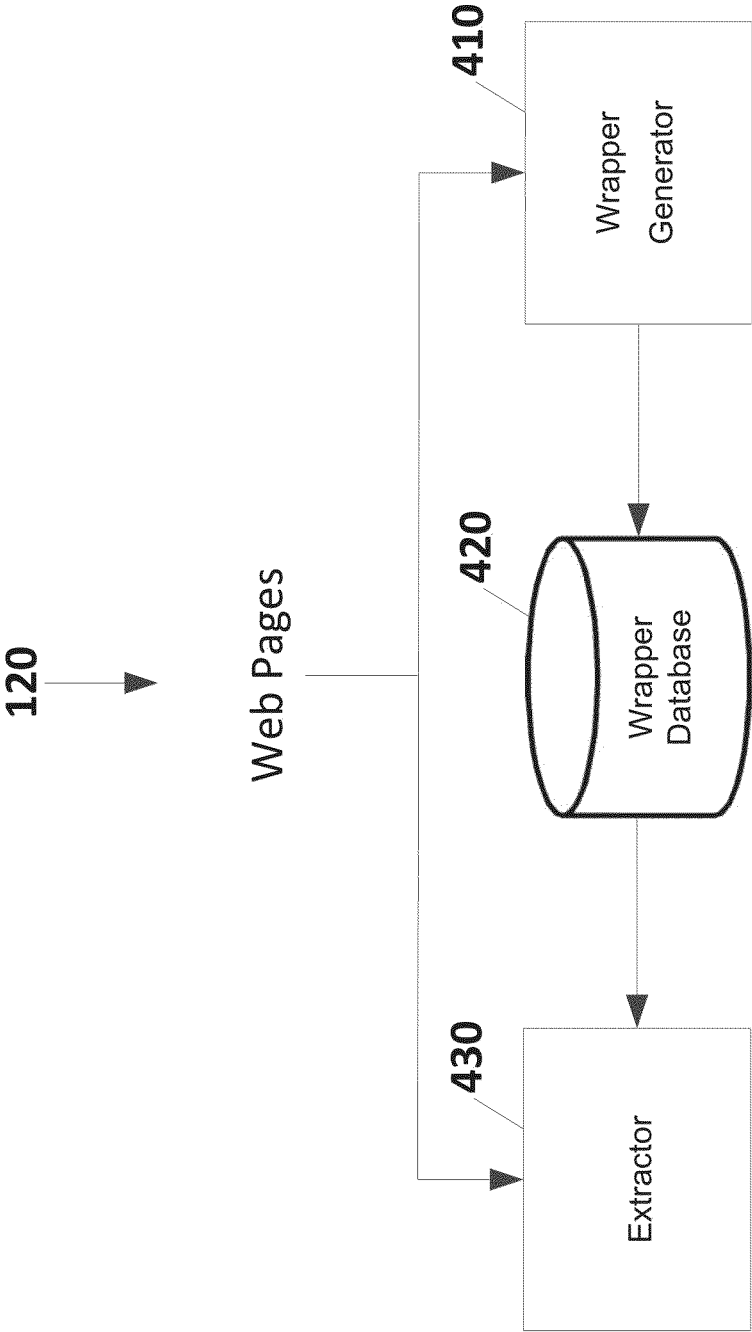


Fig. 4

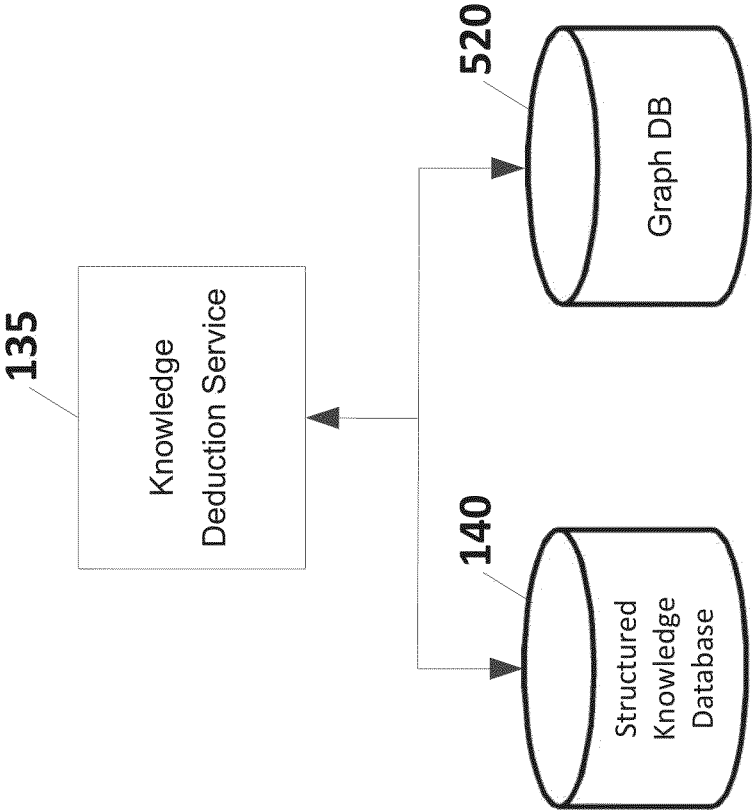


Fig. 5

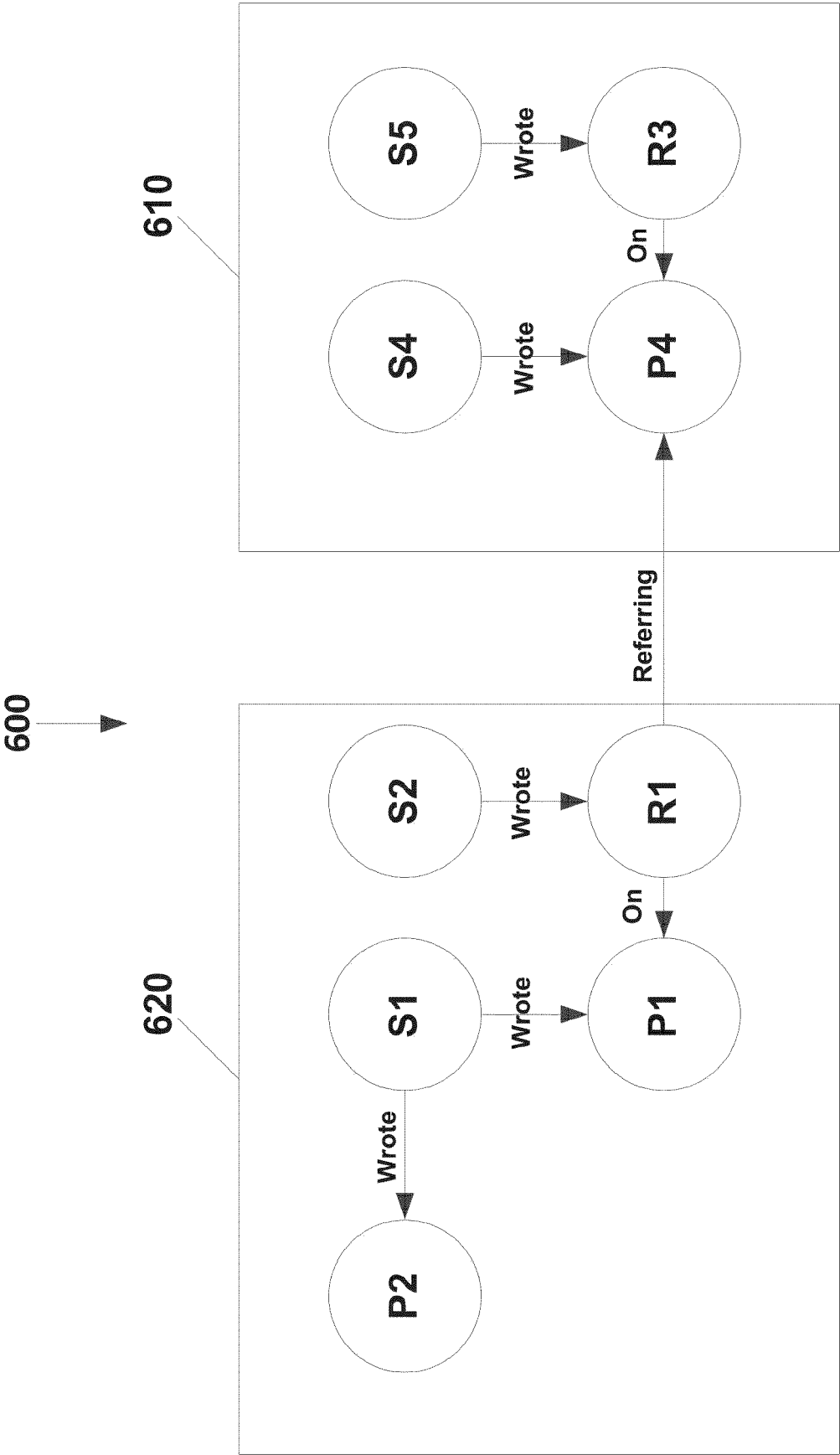


Fig. 6

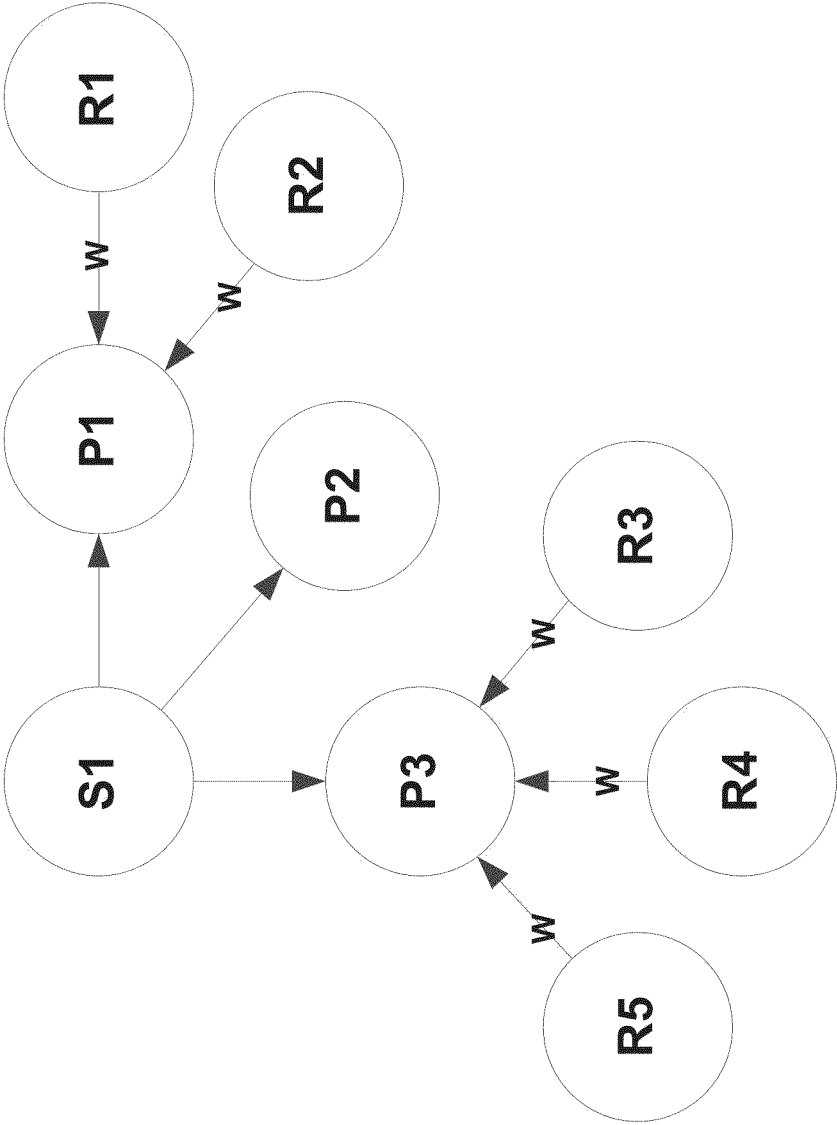


Fig. 7



800

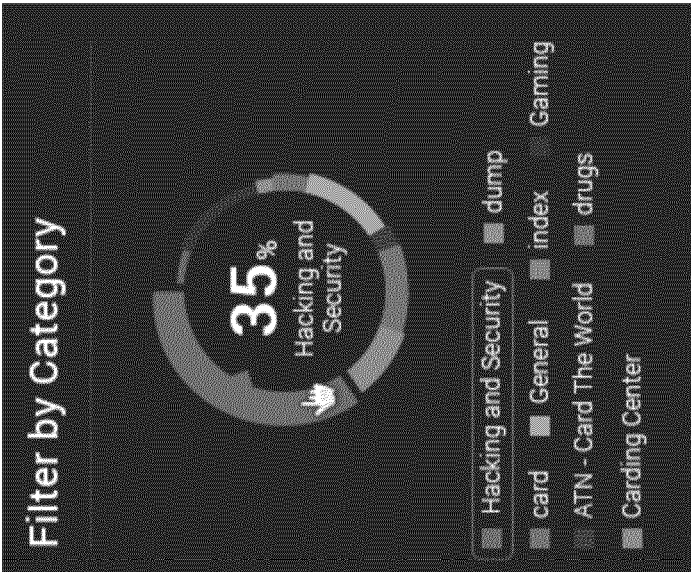


Fig. 8

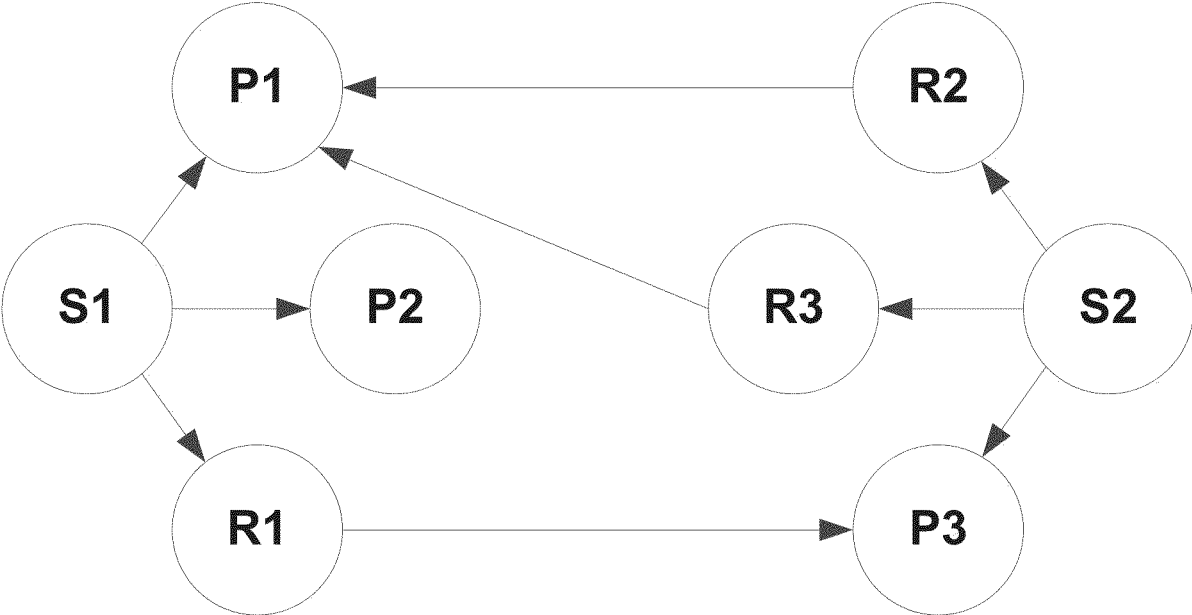


Fig. 9A

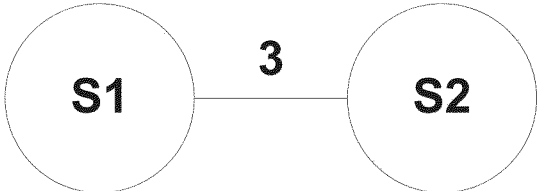


Fig. 9B

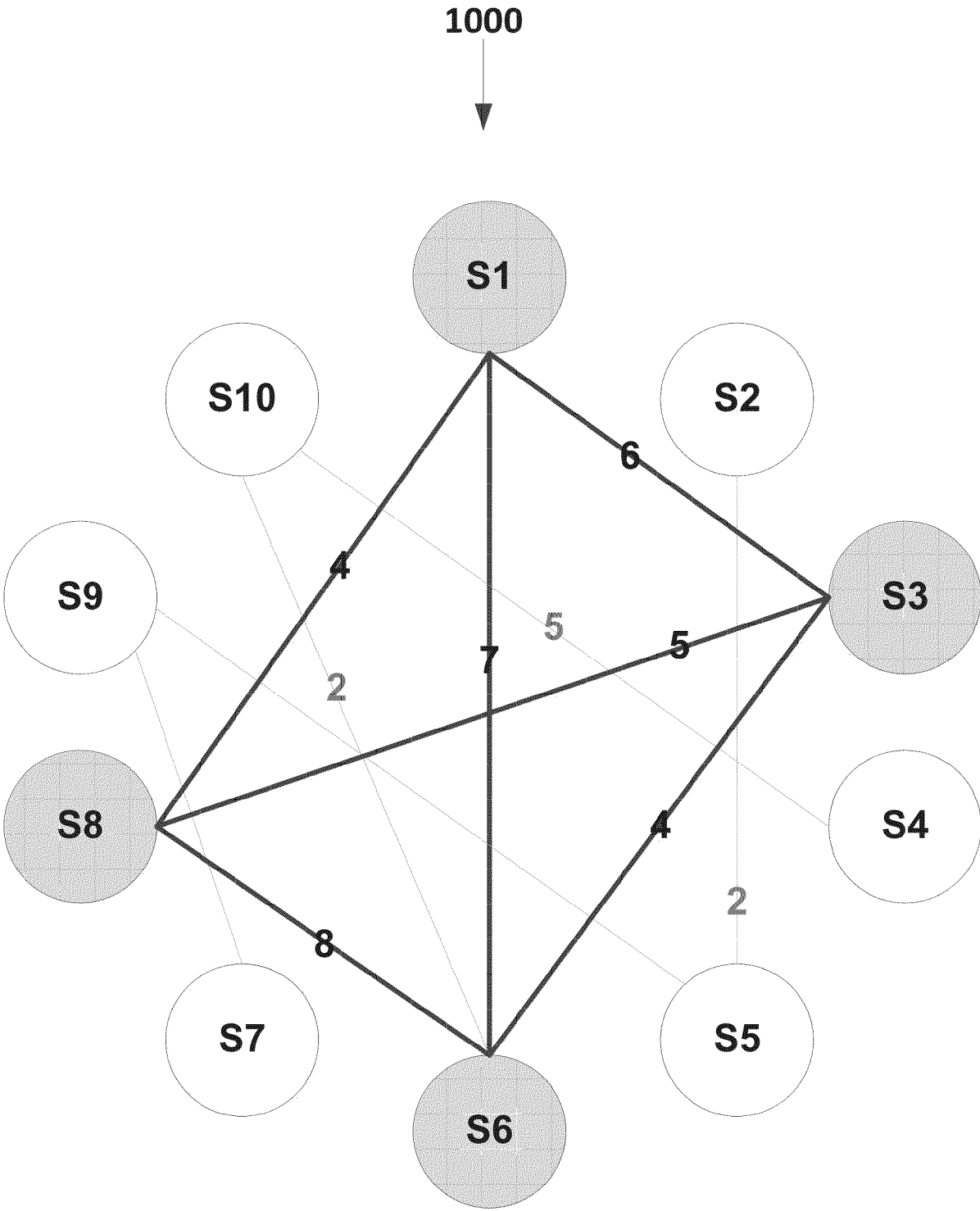


Fig. 10

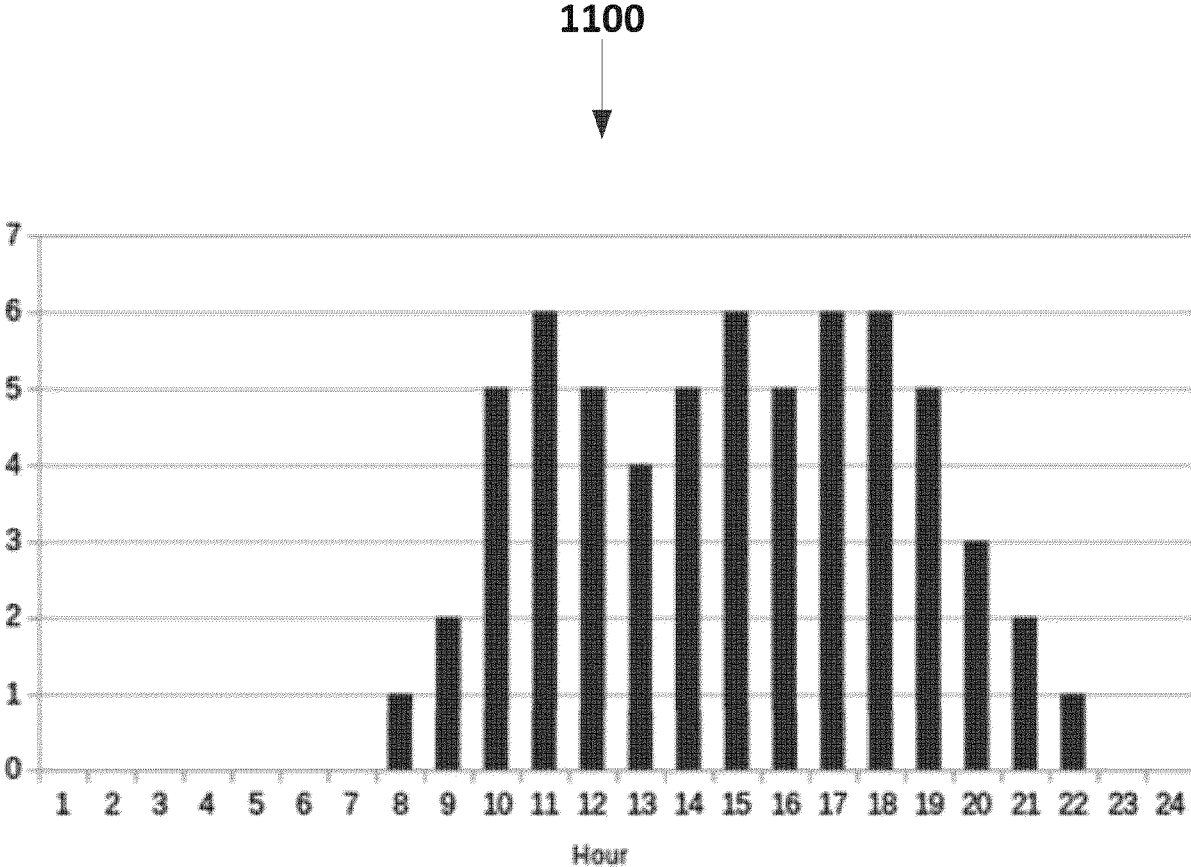



Fig. 11

1100A  


	<u>1</u>	<u>5</u>	<u>10</u>			<u>15</u>	<u>20</u>	<u>24</u>
<u>S</u>			<u>5</u>	<u>4</u>	<u>4</u>	<u>2</u>	<u>5</u>	
<u>M</u>			<u>5</u>	<u>4</u>	<u>4</u>	<u>3</u>	<u>5</u>	
<u>T</u>			<u>5</u>	<u>4</u>	<u>3</u>	<u>4</u>	<u>5</u>	
<u>W</u>		<u>5</u>	<u>5</u>	<u>4</u>	<u>1</u>	<u>2</u>	<u>5</u>	
<u>T</u>		<u>2</u>						
<u>F</u>								
<u>S</u>								

Fig. 11A

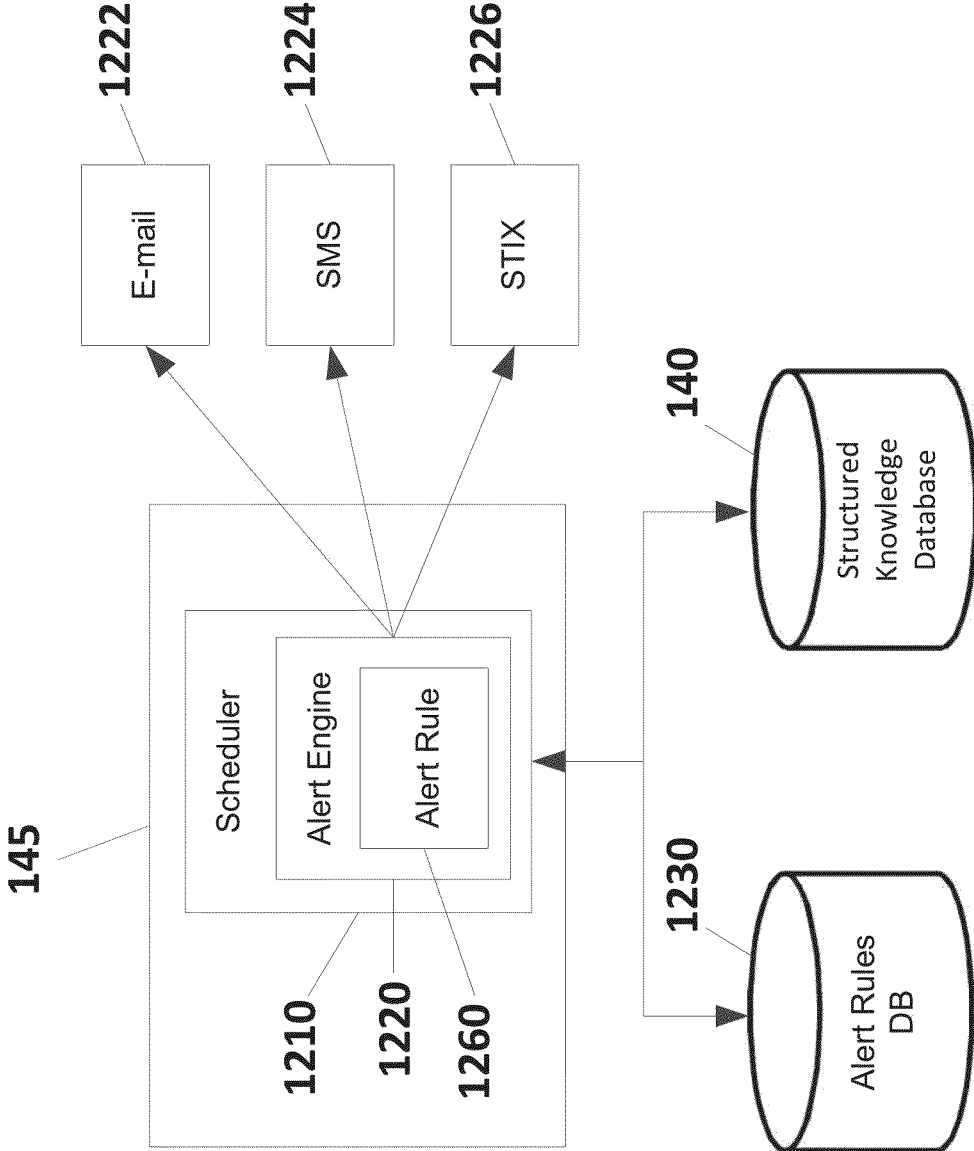


Fig. 12

1300



The image shows a 'Create Alert' dialog box with a dark background and white text. At the top right is a close button (X). The main heading is 'Notify when the following keyword sare mentioned in any post or reply'. Below this is a horizontal list of keywords: 'Card', 'NOT credit', and 'actor: Moshe', each with a small X to its right. A larger X is at the end of the list. Below the keywords is a checkbox labeled 'Include connections in actor's social network'. Underneath is the 'Alert Name:' field with the text 'Follow post'. The 'Importance:' section has a dropdown menu set to 'Very Important'. The 'Notify by:' section has two options: 'Send notification to' (checked) with a dropdown set to 'CT Group', and 'Add results to case' (unchecked) with a dropdown set to 'Select'. The 'How often:' section has three radio button options: 'Once a day at' (selected) with a time dropdown set to '24:00' and a '+ Add another time' link; 'Once a week at' with a day dropdown set to 'Monday' and a time dropdown set to '09:00'; and 'As soon as possible'.

Fig. 13

## DARK WEB MONITORING, ANALYSIS AND ALERT SYSTEM AND METHOD

### RELATED APPLICATIONS

[0001] This application is a continuation of U.S. Pat. Application No. 16/066,315 filed on Jun. 27, 2018, which is a National Phase of PCT Patent Application No. PCT/IB2016/058016 having International Filing Date of Dec. 27, 2016, which claims the benefit of priority under 35 USC §119(e) of U.S. Provisional Pat. Application No. 62/271,344 filed on Dec. 28, 2015. The contents of the above applications are all incorporated by reference as if fully set forth herein in their entirety.

### FIELD OF THE INVENTION

[0002] The present invention generally relates to web activities analysis and specifically to a Dark Web monitoring, analysis and alert system and method.

### BACKGROUND

[0003] The Dark Web is a term that refers specifically to a collection of websites that are publicly visible, but hide the IP addresses of the servers that run them. The dark web forms a small part of the Deep Web, the part of the Web not indexed by search engines. Thus they can be visited easily by any web user, but it is very difficult to work out who is behind the sites and search engines cannot find them.

[0004] The dark nets which constitute the dark web include small, friend-to-friend peer-to-peer networks, as well as large, popular networks like Freenet, I2P, and Tor, operated by public organizations and individuals. Users of the dark web refer to the regular web as the Clear net due to its unencrypted nature. The Tor dark web may be referred to as Onion land, a reference to the network's name as "the onion router."

[0005] Almost all sites on the so-called Dark Web hide their identity using, for example, the Tor encryption tool. Tor can be used to hide your identity, and spoof your location.

[0006] To visit a site on the Dark Web that is using Tor encryption, the web user needs to be using Tor. Just as the end user's IP is bounced through several layers of encryption to appear to be at another IP address on the Tor network, so is that of the website.

[0007] Because of the nature of the Dark Web and the illegal activities it enables, there is a long felt need for a system that monitors the Dark Web, analyses harvested data and provides alerts according to defined parameters.

### SUMMARY OF THE INVENTION

[0008] According to an aspect of the present invention, there is provided a dark web monitoring, analysis and alert system comprising at least one Crawler connected with the dark web, the at least one Crawler configured to scan and collect data from the dark web; a Structured Data Extractor (SDE) connected with the at least one Crawler, the SDE configured to analyze the collected data provided by the at least one Crawler and to extract structural parameters; a Structured Data Database (SDD) connected with the SDE, The SDD configured to store the structural parameters extracted by the SDE; a Text Search and Analytic Engine

(TSAE) connected with the SDD, the TSAE configured to enable advanced search and basic analysis in the collected data; a Knowledge Deduction Service (KDS) connected with the TSAE, the KDS configured to deeply analyze the collected data; a Structured Knowledge Database (SKD) connected with the KDS, the SKD configured to store the deep analysis results; and an Alert Service connected with the TSAE and the SKD, the Alert Service configured to provide alerts based on the deep analysis.

[0009] The basic analysis may comprise at least one of date in which most of comments were written, number of posts a surfer wrote for a specific search query, distribution of categories in a site, time line trending for a specific search query and top sites for a specific query.

[0010] The deep analysis may comprise at least one of finding surfers who have the highest reputation and monitoring their activities, and monitoring surfers' activity hours, social connections and group dynamics.

[0011] The search results may be prioritized according to at least one of: source scoring; recency; user reputation; record type scoring; search result relevance scoring; and content analysis scoring.

[0012] The system may further comprise a Hidden Service Locator (HSL) connected with the dark web and the at least one Crawler, the HSL may be configured to find hidden Uniform Resource Locators (URLs) in the dark web; the at least one Crawler may further be configured to scan and collect data from the dark web using the URLs.

[0013] The HSL may comprise a Tor Relay (TR) configured to blend among relays of The Onion Router (TOR) network, the TR may be configured to keep a record of URLs routed therethrough.

[0014] The at least one Crawler may comprises an IP Changer Proxy (IPCP) connected with the dark web, the IPCP configured to manage the Internet Protocol (IP) address of the at least one Crawler; a Spider connected with the IPCP, the Spider configured to progress from one web page to another; the Spider comprises a Link Extractor configured to extract URLs it finds in each web page; a Page Classifier and URL Filtering Module (PCUFM) connected with the Spider and the HSL, the PCUFM configured to classify the web pages extracted by the Spider; and a Crawler Control Center (CCC) connected with the Spider, the CCC configured to control the operation timing and the pace of data collection of the at least one Crawler.

[0015] The IP address managing may comprise at least one of: hide the at least one Crawler's real IP address and change the at least one Crawler's IP address.

[0016] The progress may be performed by extracting URLs the Link Extractor finds in each web page.

[0017] The extracted URLs may be saved in a URL Repository.

[0018] The PCUFM may further be configured to filter unwanted or unnecessary URLs and save the remaining URLs in a URL Repository.

[0019] The system may further comprise a Configuration Database configured to store Crawler's configuration.

[0020] The Crawler's configuration may comprise at least one of initial URLs for the at least one Crawler to start from, username(s) and password(s) of the at least one Crawler and the at least one Crawler's timing setting.

[0021] The system may further comprise a Web Content Cache configured to store web pages extracted by the Spider.



**[0022]** The at least one Crawler may further be configured to optimize its scanning pace versus its secrecy.

**[0023]** The optimizing may comprise decreasing the scanning pace and changing the at least one Crawler's IP address.

**[0024]** The optimizing may comprises changing the at least one Crawler's username.

**[0025]** The Structured Data Extractor may comprises a Wrapper Generator; a Wrapper Database; and an Extractor; the Wrapper Generator may be configured to analyze a web page, find patterns, create a wrapper and save the wrapper in the Wrapper Database; the Extractor may be configured to receive a web page and a suitable wrapper from the Wrapper Database and to extract relevant data from the page according the wrapper.

**[0026]** The wrapper may comprises labels.

**[0027]** The Knowledge Deduction Service may further be configured to classify posts into categories and to analyze the sentiment of comments.

**[0028]** The sentiments may comprise negative, positive and neutral sentiments.

**[0029]** The Text Search and Analytic Engine may further be configured to determine a surfer's fields of interest by summing the surfer's posts in each category.

**[0030]** The Knowledge Deduction Service may further be configured to identify groups by monitoring the number of interactions between surfers.

**[0031]** The Knowledge Deduction Service may further be configured to perform an activity times analysis.

**[0032]** The activity times analysis may comprises calculating a temporal data distribution within a time frame; saving the time frame which includes most of the data; and saving the average and the standard deviation of the temporal data distribution.

**[0033]** The Knowledge Deduction Service may further be configured to find surfers who use different aliases.

**[0034]** The finding may comprises at least one of locating communication information used by more than one surfer; looking for similar aliases excluding common names; locating surfers with similar activity pattern using the activity times analysis; locating surfers with similar fields of interest; locating surfers who are active for a certain period and then continue the activity in other places/other aliases; locating surfers who post the same content at the same time in two different locations; counting the most frequent words used by a surfer; and analyzing surfers' text.

**[0035]** The analyzing surfers' text may comprise at least one of the use of punctuation marks, upper/lower case and common misspelling.

**[0036]** The Alert Service may comprise a Scheduler configured to schedule the monitoring related to each alert; an Alert Engine configured to send alerts; and an Alert Rule module.

**[0037]** The alerts may be sent via at least one of e-mail and Short Message Service (SMS).

**[0038]** The alerts may be sent according to rules written in an Alert Rules Database and prioritized according to processed data stored in the Structured Knowledge Database.

**[0039]** The prioritization may comprises at least one of: source scoring; recency; user reputation; record type scoring; search result relevance scoring; and content analysis scoring.

**[0040]** The Alert Service may comprise: a Scheduler configured to schedule the monitoring related to each alert; an Alert Engine configured to send alerts; and an Alert Rule.

**[0041]** The alerts may be sent according to rules written in an Alert Rules Database and prioritized according to the prioritized search results.

**[0042]** The Alert Rule module may be configured to at least one of: define wake up intervals, enable search by a key word, enable search by an activity related to a certain surfer, enable search by an activity of a certain group, enable search by a change in trend of a certain key word and enable search by a new phrase or a word that appears more than a predetermined number of times.

**[0043]** The system may further comprise a case management module configured to enable a client of the system to create a case file in order to manage a research or an investigation.

**[0044]** The system may further comprise a recommendation engine configured to recommend adding relevant surfers and/or posts to the case file.

**[0045]** The recommendation may be performed according to at least one of: building a connection map of existing surfers in the case file, analyzing the connections and recommending adding surfers that have a strong connection with the existing surfers in the case file; "similar" surfers; surfers that published posts collected in the case file; surfers that are mentioned in existing posts' content; surfers having similar fields of interest; and posts that have a strong contextual matching.

**[0046]** The contextual matching may comprise at least one of: same classification, same time in the time range of posts in the case file and posts having a words-matching up to a certain threshold.

**[0047]** According to another aspect of the present invention, there is provided a method of dark web monitoring, analyzing and providing alerts, comprising: receiving client's preferences for defining at least one alert; providing data collected from the dark web and structured; performing an advanced search and basic analysis in the structured data based on the client's preferences; performing deep analysis of the structured data based on the client's preferences; and providing the at least one alert based on the deep analysis.

**[0048]** Providing the data may comprise scanning and collecting data from the dark web by at least one Crawler, analyzing the collected data provided by the at least one Crawler and extracting structural parameters.

**[0049]** The method may further comprise storing the structural parameters.

**[0050]** The basic analysis may comprise the date in which most of comments were written, how many posts a surfer wrote for a specific search query, distribution of categories in a site, time line trending for a specific search query and top sites for a specific query.

**[0051]** The search results may be prioritized according to at least one of: source scoring; recency; user reputation; record type scoring; search result relevance scoring; and content analysis scoring.

**[0052]** The deep analysis may comprise at least one of finding surfers who have the highest reputation and monitoring their activities, and monitoring surfers' activity hours, social connections and group dynamics.

**[0053]** The method may further comprise storing the deep analysis results.

**[0054]** The method may further comprise finding hidden Uniform Resource Locators (URLs) in the dark web and scanning and collecting data from the dark web using the hidden URLs.

**[0055]** The method may further comprise blending by a Tor Relay (TR) among relays of The Onion Router (TOR) network and keeping a record of URLs routed through the TOR.

**[0056]** The at least one Crawler may comprise: managing Internet Protocol (IP) address of the at least one Crawler; progressing from one web page to another and extracting URLs found in each web page; classifying the extracted web pages; and controlling operation timing and pace of data collection.

**[0057]** The IP address managing may comprise at least one of: hiding the at least one Crawler's real IP address and changing the at least one Crawler's IP address.

**[0058]** The progressing may comprise extracting URLs found in each web page.

**[0059]** The method may further comprise saving the extracted URLs.

**[0060]** The method may further comprise filtering unwanted or unnecessary URLs and saving the remaining URLs.

**[0061]** The method may further comprise storing the at least one Crawler's configuration.

**[0062]** The at least one Crawler's configuration may comprise at least one of initial URLs for the at least one Crawler to start from, username(s) and password(s) of the at least one Crawler and the at least one Crawler's timing setting.

**[0063]** The method may further comprise storing extracted web pages.

**[0064]** The at least one Crawler may further comprise optimizing its scanning pace versus its secrecy.

**[0065]** The optimizing may comprise decreasing the scanning pace and changing the at least one Crawler's IP address.

**[0066]** The optimizing may comprise changing the at least one Crawler's username.

**[0067]** The method may further comprise: analyzing a web page, finding patterns, creating a wrapper and saving the wrapper; and receiving a web page and a suitable wrapper and extracting relevant data from the page according to the wrapper.

**[0068]** The wrapper may comprise labels.

**[0069]** The method may further comprise classifying posts into categories and analyzing the sentiment of comments.

**[0070]** The sentiments may comprise negative, positive and neutral sentiments.

**[0071]** The method may further comprise determining a surfer's fields of interest by summing the number of the surfer's posts in each category.

**[0072]** The method may further comprise identifying groups by monitoring the number of interactions between surfers.

**[0073]** The method may further comprise performing an activity times analysis.

**[0074]** The activity times analysis may comprise: calculating a temporal data distribution within a time frame; saving the time frame which includes most of the data; and saving the average and the standard deviation of the temporal data distribution.

**[0075]** The method may further comprise finding surfers who use different aliases.

**[0076]** Finding surfers may comprise at least one of: locating communication information used by more than one surfer; looking for similar aliases excluding common names; locating surfers with similar activity pattern using the activity times analysis; locating surfers with similar fields of interest; locating surfers who are active for a certain period and then continue the activity in other places/other aliases; locating surfers who post the same content at the same time in two different locations; counting the most frequent words used by a surfer; and analyzing surfers' text.

**[0077]** Analyzing surfers' text may comprise at least one of the use of punctuation marks, upper/lower case and common misspelling.

**[0078]** The providing the at least one alert may comprise: scheduling the monitoring related to each alert; and sending alerts.

**[0079]** The method may further comprise sending the alerts via at least one of e-mail and Short Message Service (SMS).

**[0080]** The method may further comprise sending the alerts according to rules and prioritized according to processed data.

**[0081]** The prioritization may comprise at least one of: source scoring; recency; user reputation; record type scoring; search result relevance scoring; and content analysis scoring.

**[0082]** The providing the at least one alert may comprise: scheduling the monitoring related to each alert; and sending alerts.

**[0083]** The method may further comprise sending the alerts according to rules written in an Alert Rules Database and prioritized according to the prioritized search.

**[0084]** The method may further comprise at least one of: defining wake up intervals, enabling search by a key word, enabling search by an activity related to a certain surfer, enabling search by an activity of a certain group, enabling search by a change in trend of a certain key word and enabling search by a new phrase or a word that appears more than a predetermined number of times.

**[0085]** The method may further comprise creating a case file in order to manage a research or an investigation.

**[0086]** The method may further comprise providing recommendations for adding relevant surfers and/or posts to the case file.

**[0087]** The recommendation may be performed according to at least one of: building a connection map of existing surfers in the case file, analyzing the connections and recommending adding surfers that have a strong connection with the existing surfers in the case file; "similar" surfers; surfers that published posts collected in the case file; surfers that are mentioned in existing posts' content; surfers having similar fields of interest; and posts that have a strong contextual matching.

**[0088]** The contextual matching may comprise at least one of: same classification, same time in the range of posts in the case file and posts having a words-matching up to a certain threshold.

**[0089]** According to another aspect of the present invention, there is provided a dark web monitoring, analysis and alert system comprising: a data receiving module configured to receive data collected from the dark web and structured; a Structured Data Database (SDD) connected with the data receiving module, the SDD configured to store the structured data; a Text Search and Analytic Engine (TSAE) con-

nected with the SDD, the TSAE configured to enable advanced search and basic analysis in the structured data; a Knowledge Deduction Service (KDS) connected with the TSAE, the KDS configured to deeply analyze the collected data; the deep analysis comprises extracting insights regarding dark web surfers behavioral patterns and interactions; a Structured Knowledge Database (SKD) connected with the KDS, the SKD configured to store the deep analysis results; and an Alert Service connected with the TSAE and the SKD, the Alert Service configured to provide prioritized alerts based on the deep analysis.

#### BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

**[0090]** For better understanding of the invention and to show how the same may be carried into effect, reference will now be made, purely by way of example, to the accompanying drawings.

**[0091]** With specific reference now to the drawings in detail, it is stressed that the particulars shown are by way of example and for purposes of illustrative discussion of the preferred embodiments of the present invention only, and are presented in the cause of providing what is believed to be the most useful and readily understood description of the principles and conceptual aspects of the invention. In this regard, no attempt is made to show structural details of the invention in more detail than is necessary for a fundamental understanding of the invention, the description taken with the drawings making apparent to those skilled in the art how the several forms of the invention may be embodied in practice. In the accompanying drawings:

**[0092]** FIG. 1 is a schematic view of the system according to embodiments of the present invention;

**[0093]** FIG. 2 is a schematic view of the Crawler of FIG. 1 according to embodiments of the present invention;

**[0094]** FIG. 3 is a flowchart showing the process performed by the Crawler according to embodiments of the present invention;

**[0095]** FIG. 4 is a schematic view of the Structured Data Extractor of FIG. 1 according to embodiments of the present invention;

**[0096]** FIG. 5 is a schematic view of the data sources connected with the Knowledge Deduction Service of FIG. 1;

**[0097]** FIG. 6 represents a graph of two forums connected via a "Referring" from one forum to the other;

**[0098]** FIG. 7 shows an exemplary surfer and his posts and comments;

**[0099]** FIG. 8 shows an exemplary representation of the fields of interest of an exemplary surfer;

**[0100]** FIG. 9A shows an exemplary graph representing the interactions between two surfers;

**[0101]** FIG. 9B shows an exemplary graph representing the total number of interactions between the two surfers of FIG. 9A;

**[0102]** FIG. 10 shows an exemplary interactions graph, where the groups that have at least four interactions are highlighted;

**[0103]** FIG. 11 shows an exemplary data distribution graph represented by the number of activities in each hour.

**[0104]** FIG. 11A shows another exemplary data distribution graph using a matrix of 24X7;

**[0105]** FIG. 12 is a schematic view of the Alert Service of FIG. 1 components according to embodiments of the present invention; and

**[0106]** FIG. 13 shows an exemplary user interface for creating an alert according to embodiments of the present invention.

#### DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

**[0107]** Before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and the arrangement of the components set forth in the following description or illustrated in the drawings. The invention is applicable to other embodiments or of being practiced or carried out in various ways. Also, it is to be understood that the phraseology and terminology employed herein is for the purpose of description and should not be regarded as limiting.

**[0108]** The present invention provides a dark web monitoring, analysis and alert system enabling to monitor dark web activities taking place in online stores, forums, etc. and provide information and alerts when suspicious threats are detected.

**[0109]** The Dark Web is a term that refers to a collection of websites that are publicly visible, but hide the IP addresses of the servers that run them. The dark web forms a small part of the Deep Web, the part of the web not indexed by search engines. Thus they can be visited by any web user, but it is very difficult to work out who is behind the sites.

**[0110]** It will be appreciated that the term Dark Web here and below may refer to any part of the web including any part of the Deep Web and preferably the part of the web in which surfers are anonymous.

**[0111]** The system of the present invention tracks and monitors anonymous surfers, analyzes their activities and their social network thus enabling to track surfers even if they change their identity.

**[0112]** FIG. 1 is a schematic view of the system 100 according to embodiments of the present invention, comprising:

**[0113]** at least one Crawler 110 that scans and collects information from the dark web 105 and other relevant web sites (optional);

**[0114]** a Hidden Service Locator 115 that finds hidden URLs (optional);

**[0115]** a Structured Data Extractor 120 that analyses html pages and extracts structural parameters (optional);

**[0116]** a Structured Data Database 125 that stores the structural parameter extracted by the Structured Data Extractor 120;

**[0117]** a Text Search and Analytic Engine 130 that enables advanced search in the Structured Data Database 125 and basic analysis;

**[0118]** a Knowledge Deduction Service 135 that deeply analyses the data;

**[0119]** a Structured Knowledge Database 140;

**[0120]** an Alert Service 145 that provides alerts.

**[0121]** According to embodiments of the invention, the system of the present invention may comprise only the Structured Data Database 125, the Text Search and Analytic Engine 130, the Knowledge Deduction Service 135, the Structured Knowledge Database 140 and the Alert Service 145. The URLs, the data and/or the structured data may be provided to the system by a data provider, received via a

data receiving module and stored in the Structured Data Database **125**.

**[0122]** The uniqueness of the Crawler(s) **110** of the present invention is the ability to:

**[0123]** 1. Deal with systems that detect Crawlers, namely, disguise as a “regular” surfer.

**[0124]** 2. Control the timing and amount of data collection.

**[0125]** 3. Change its own IP addresses.

**[0126]** It will be appreciated that the Crawler **110** is not limited to these exemplary abilities. Alternatively, it may have at least one of these abilities or more than these three described.

**[0127]** FIG. 2 is a schematic view of the Crawler **110** of FIG. 1 according to embodiments of the present invention, comprising:

**[0128]** an IP Changer Proxy **210** that manages the IP address of the Crawler **110** in order to hide the Crawler’s real address and changes the Crawler’s address in cases where the Crawler logs in to a web site with different usernames; the IP Changer Proxy also mediates between the Crawler **110** protocol and internet protocols (e.g. internet relay chat (IRC), Hypertext Transfer Protocol (http), etc.)

**[0129]** a Spider **215** that progresses from one web page to another by extracting the URLs it finds in each web page. The Spider **215** comprises a Link Extractor **220** that extracts and saves the URLs it finds in each web page in a URL Repository **245**;

**[0130]** a Page Classifier and URL Filtering module **225** that classifies the web pages extracted by the Spider **215**, filters the unwanted or unnecessary URLs (e.g. URLs from Google) and saves the remaining URLs in the URL Repository **245**;

**[0131]** a Crawler Control Center **230** that controls the Crawler’s operation timing and the pace of data collection;

**[0132]** a Configuration Database **235** that stores the required Crawler configuration, for example, initial URLs for the Crawler to start from, username(s) and password(s) of the Crawler and the Crawler’s timing setting (if changed from default), e.g. the number of requests per day, the number of samples per day, etc. According to embodiments of the invention, the initial URLs may be provided manually by an analyst. Alternatively or additionally, the initial URLs may be extracted from the internet (e.g. from Google).

**[0133]** a Web Content Cache **240** that stores web pages extracted by the Spider **215**.

**[0134]** FIG. 3 is a flowchart **300** showing the process performed by the Crawler **110** according to embodiments of the present invention. In step **310**, the Spider **210** visits and reads URLs out of the URL Repository **245** and updates the last visit date for each URL. Reading URLs from the same domain is random. The random reading assists in being undetectable. In step **320**, the Spider **210** reads the web pages using IP addresses provided by the IP Changer Proxy **210** in order to hide its real IP address. In step **330**, for each web page, using the Link Extractor **220**, the Spider **210** extracts the URLs it finds and saves the web page in the Web Content Cache **240** in order to enable the system to process or reprocess web pages in a later phase. In step **340**, using the Page Classifier and URL Filtering module **225**, each URL from a new domain is classified and only the wanted

or needed URLs are saved in the URL Repository **245**. Wanted or needed URLs are URLs that contain a key word(s) or a topic(s) that the system is currently interested in. In step **350**, for each new domain, the Page Classifier and URL Filtering module **225** stores in the Configuration DB **235** various parameters, such as for example the web page’s header, in order to be able to locate and scan this domain in the future in case the address will be changed.

**[0135]** Many web sites have Crawler detection mechanisms. In order to avoid being detected by these mechanisms the system of the present invention:

**[0136]** 1. Controls the scanning pace - the number of scanned pages per time unit (minute, second).

**[0137]** 2. Controls the scanning duration and sequence - for example, scans for 8 hours and rests for 4 hours.

**[0138]** 3. Scans randomly - the scanning order of a web page and the extraction of web pages it contains are random.

**[0139]** 4. Scans in parallel - a plurality of Crawlers, having a plurality of IP addresses, may scan the same web page simultaneously.

#### Optimization of Scanning Pace Versus Secrecy

**[0140]** In order to be able to scan web pages which update frequently, the Crawler has to optimize its scanning pace. The Crawler may start the scan with default parameters and optimize the process during the scan. If the web page is updated frequently, faster than the Crawler is able to scan, the Crawler may, for example, increase the scanning pace and/or add another Crawler to scan with it simultaneously. If the Crawler is blocked by blocking its IP address, it may, for example, change its IP address and decrease the scanning pace. If the Crawler is blocked by banning its username (in sites that require user registration), the Crawler may, for example, replace its username, change its IP address and decrease the scanning pace.

**[0141]** In the dark web the URLs are hidden hence the challenge is to find URLs other than the ones existing in blogs, forums, etc. Moreover, some of these URLs exist only for a short term. The Hidden Service Locator (HSL) **115** of the present invention is configured to find these hidden URLs. In The Onion Router (TOR) network, for example, TOR relays enable anonymous surfing by multi stage encryption between the relays (nodes). A TOR Relay (TR) of the HSL **115** blends among the relays of the TOR network, as described for example in “Trawling for Tor Hidden Services: Detection, Measurement, Deanonimization” by Alex Biryukov, Ivan Pustogarov, Ralf-Philipp Weinmann from the University of Luxembourg (<http://www.ieee-security.org/TC/SP2013/papers/4977a080.pdf>). When URLs are routed through the TR, it keeps a record of them. These URLs are forwarded to the Crawler **110** via the Page Classifier and URL Filtering module **225** and the URL Repository **245**.

**[0142]** The Structured Data Extractor **120** analyses html pages and extracts structural parameters such as dates, posts, comments, etc. These parameters assist in building a connection map and analyzing the data.

**[0143]** FIG. 4 is a schematic view of the Structured Data Extractor **120** of FIG. 1 according to embodiments of the present invention, comprising: a Wrapper Generator **410** such as described for example in <http://www.aclweb.org/anthology/Y11-1010>. The Wrapper Generator is connected

with a Wrapper Database **420** which is connected with an Extractor **430**. When a web page arrives from a certain domain (via the Crawler **110**), if the domain is unrecognized, the page is forwarded to the Wrapper Generator **410**. The Wrapper Generator **410** analyses the web page, finds patterns, creates a wrapper and saves it in the Wrapper DB **420**. If the domain is recognized, the web page is forwarded to the Extractor **430**. The Extractor **430** receives a web page and a suitable wrapper from the Wrapper DB and extracts the relevant data from the page according to labels defined by the wrapper, for example, a label which represents a date field.

**[0144]** The Text Search and Analytic Engine **130** enables advanced search in the Structured Data Database **125** and basic analysis such as, for example, in which date most of the comments were written, how many posts a surfer wrote for a specific search query, distribution of categories in a site, time line trending for a specific search query, top sites for a specific query, etc.

**[0145]** According to embodiments of the present invention, the system of the present invention may enable a client to receive prioritized search results. The results prioritization process, calculates the score of each search result based on the following criteria:

**[0146]** 1. Source scoring - each source in the system gets a score based on the activity in the source and the value of the information it contains.

**[0147]** 2. Recency - when was the information published (two days ago, two weeks ago, one year ago, etc.).

**[0148]** 3. User reputation described below.

**[0149]** 4. Record type scoring - for example, a post in a forum gets different score than a product in a market.

**[0150]** 5. Search results relevance scoring as described, for example, in <https://www.elastic.co/guide/en/elastic-search/guide/current/scoring-theory.html>.

**[0151]** 6. Content analysis scoring - analyzing text in order to determine whether it is a code, a single word, free language, etc. where free language receives a higher score.

**[0152]** The Knowledge Deduction Service **135** deeply analyses the data, namely, extracts insights regarding dark web surfers behavioral patterns and interactions. For example, finds the surfers who have the highest reputation and monitors their activities; monitors surfers' activity hours, social connections, group dynamics, etc. Using the Knowledge Deduction Service **135** it is possible to provide alerts, built from various pieces of data which are not necessarily directly connected to each other.

**[0153]** FIG. 5 is a schematic view of the data sources connected with the Knowledge Deduction Service **135** of FIG. 1, comprising the Structured Knowledge Database **140** and a Graph DB **520**. The Knowledge Database **140** stores information which was concluded during the data analysis (analysis results). The Graph DB **520** stores the connections between entities.

#### Reputation Evaluation

**[0154]** FIG. 6 represents a graph **600** of two forums, **610** and **620**, connected via a "Referring" from one forum (**620**) to the other (**610**). In the circles (nodes): S represents a surfer, P represents a post and R represents a response. On the lines between the circles (edges): Wrote represents - a surfer who writes a post or a comment, On represents- a surfer who

responds to a post and Referring represents- a post or a comment which refers to another post or a comment.

**[0155]** Prior to the reputation evaluation process, the Knowledge Deduction Service **135**:

**[0156]** 1. Classifies each post to its relevant category, for example, Hacking, Programming, Carding, Anonymity, etc. The classification and categorization is based on standard methods such as for example, Support Vector Machine (SVM), Bayesian, Neural Network, etc.

**[0157]** 2. Analyzes the comments and determines the sentiment value of each comment. The sentiment value ranges from -1 to +1, where +1 represents positive sentiment, -1 represents negative sentiment and 0 represents neutral sentiment. The determination may be done based on statistical calculations, on NLP (Natural Language Programming) methods and the like.

**[0158]** FIG. 7 shows an exemplary surfer S1 who wrote three posts (P1, P2 and P3) where P1 has two responses and P3 has three responses. W represents the sentiment of each response.

**[0159]** According to embodiments of the invention, the reputation evaluation is performed according to the exemplary following formulas: When

$$Qp \neq \emptyset: R(p) = \left( \{S(q) | q \text{ in } Qp\} * \sqrt[4]{|Qp|} \right)$$

When

$$Qp = \emptyset: R(p) = R0$$

When

$$Pu \neq \emptyset: R(u) = \left( \{R(p) | p \text{ in } Pu\} * \sqrt[4]{|Pu|} \right)$$

When

$$Pu = \emptyset: R(u) = 0$$

where:

**[0160]** u = user

**[0161]** p = post

**[0162]** Pu = user post list

**[0163]** Qp = post comments list, not including the user comments on its own post.

**[0164]** R(p) = post reputation

**[0165]** R(u) = user reputation

**[0166]** S(q) = sentiment of comment q (where  $-1 \leq S(q) \leq 1$ )

**[0167]** R0 = reputation of a post with no comments

**[0168]** (G) = the average value of G

**[0169]** |G| = the number of members in G

#### Fields of Interest

**[0170]** In order to monitor a surfer's fields of interest the Text Search & Analytic Engine **130** summarizes the number of surfer's posts in each category.

**[0171]** FIG. 8 shows an exemplary representation **800** of the fields of interest of an exemplary surfer.

### Group Identification

[0172] FIG. 9A shows an exemplary graph representing the interactions between surfer S1 and surfer S2.

[0173] FIG. 9B shows an exemplary graph representing the total number of interactions between surfer S1 and surfer S2.

[0174] FIG. 10 shows an exemplary interactions graph 1000 where the groups that have at least four interactions are highlighted.

### Activity Times Analysis

[0175] The activity times analysis enables the system to monitor the behavior of surfers, whether they are “full time” surfers, amateurs, night surfers, etc. Moreover, it may provide indication on the location of the surfers.

[0176] In order to perform the activity times analysis the Knowledge Deduction Service 135:

[0177] 1. Calculates the temporal data distribution within a time frame of e.g. 24 hours or 7 days.

[0178] 2. Saves the time frame which includes most of the data.

[0179] 3. Saves the average and the standard deviation of the temporal data distribution.

[0180] FIG. 11 shows an exemplary data distribution graph 1100 represented by the number of activities (e.g. posts) in each hour.

[0181] FIG. 11A shows another exemplary data distribution graph 1100A using a matrix of 24X7 (hours X days). Each cell represents the sum of the number of records (in term of the hour and day), for example, on Sunday at 10AM four posts were written. The data is normalized to a range of 0-5 by, for example, the following formula:

$$\text{Ceil}(\log_4(x+1))$$

where X is the sum of records in one cell.

### Surfer's Identity Matching

[0182] The surfer's identity matching process enables to find surfers who use different aliases (nicknames).

[0183] In order to find such surfers the Knowledge Deduction Service 135 may:

[0184] 1. Locate communication information (e.g. email, ICQ, jabber, etc) and locate other identities that are using the same communication information. To achieve that, the system detects a false matching (such as a reference to communication information by another surfer) by using machine learning techniques.

[0185] 2. Look for similar aliases excluding common names (e.g. guest, anonymous).

[0186] 3. Locate surfers with similar activity pattern using the activity times analysis.

[0187] 4. Locate surfers with similar fields of interest.

[0188] 5. Locate surfers who are active for a certain period and then continue the activity in other places/other aliases. For example, a surfer was active from the first of November until the end of the month and then active in another place/with other aliases from the first of December.

[0189] 6. Locate surfers who post the same content at the same time in two different locations.

[0190] 7. Count the most frequent words used by a surfer (excluding stop words and other common words).

[0191] 8. Analyze the surfer's text, the use of punctuation marks, upper/lower case, common misspelling, etc.

### Surfers Profile Characterization

[0192] The surfer profile characterization is the adjusted calculation of the reputation evaluation, the fields of interest monitoring, the group identification and the activity times analysis of the surfer across multiple sources and aliases.

[0193] The Alerts Service 145 of the present invention is a unique tool providing prioritized alerts to clients who use the system of the present invention via various media such as e-mail, Short Message Service (SMS), standard cyber threat intelligent format (STIX), etc. The alerts may be generated based on the client's preferences, for example, when a certain word, in a certain field, written by a certain surfer is monitored. The system of the present invention may be integrated in the client's alerts system in order to strengthen the client's alerts capabilities.

[0194] FIG. 12 is a schematic view of the Alert Service 145 of FIG. 1 components according to embodiments of the present invention, comprising a Scheduler 1210 which Schedules the monitoring process related to each alert, an Alert Engine 1220 which sends alerts (e.g. via e-mail 1222, SMS 1224 and STIX 1226) according to rules written in the Alert Rules DB 1230 and prioritizes them according to the processed data stored in the Structured Knowledge Database 140; and an Alert Rule module 1260.

[0195] The Scheduler 1210 wakes up once in a while according to predetermined time periods configured in the Alert Rules DB 1230. The Alert Engine 1220 analyzes the rules it has to perform according to the rules written in the Alert Rules DB 1230, scans the data stored in the Text Search and Analytic Engine (130 of FIG. 1) and the Structured Knowledge Database 140 accordingly, prioritizes the alerts as described below and sends alerts accordingly.

[0196] The Alert Rule module 1260 defines the wake up intervals, enables search by a key word, an activity related to a certain surfer, an activity of a certain group, a change in trend of a certain key word, a new phrase or a word that appears more than a predetermined number of times, a combination of the above, etc.

[0197] FIG. 13 shows an exemplary user interface 1300 for creating an alert. According to embodiments of the invention the user may define key words, the name of the alert, the importance level of the alert, the addressees of the alert and the frequency of alerting.

[0198] It will be appreciated that the present invention is not limited to the above exemplary definitions.

### Alerts Prioritization

[0199] According to embodiments of the invention, an alert prioritization process, performed by the Alert Engine, may calculate the score of each alert based on the following criteria:

[0200] 1. Source scoring - each source in the system gets a score based on the activity in the source and the value of the information it contains.

[0201] 2. Recency - when was the information published (two days ago, two weeks ago, one year ago, etc.).

[0202] 3. User reputation described above.

[0203] 4. Record type scoring - for example, a post in a forum gets different score than a product in a market.

[0204] 5. Search results relevance scoring as described, for example, in <https://www.elastic.co/guide/en/elastic-search/guide/current/scoring-theory.html>.

[0205] 6. Content analysis scoring - analyzing text in order to determine whether it is a code, a single word, a free language, etc. where free language receives a higher score.

[0206] Alternatively or additionally, the alert prioritization process may use the results of the search prioritization process described above.

[0207] An Analytical Dashboard of the present invention enables to view the data analysis described above. The Analytical Dashboard may comprise categories, number of posts by dates, search results, an option to create an alert from this search, the total number of search results, etc.

[0208] A surfer Analytic Dashboard of the present invention enables to view data analysis of a certain surfer, comprising the surfer's details, his activity analysis, number of posts by dates, categories, his connection map, etc.

[0209] According to embodiments of the present invention, the system of the present invention may further comprise a case management module enabling a client of the system to create a case file in order to manage a research or an investigation by adding posts, surfers and alerts notifications.

[0210] According to embodiments of the present invention, the system 100 may further comprise a recommendation engine which may recommend adding relevant surfers, posts, etc. to the case file, for example, by building a connection map of the existing surfers in the case file, analyzing the connections and recommending adding surfers that have a strong connection with the existing surfers in the case file.

[0211] According to embodiments of the present invention, the recommendation engine may recommend adding to the case file:

[0212] 1. "Similar" surfers based on the Surfer's identity matching described above.

[0213] 2. Surfers that published the posts collected in the case file.

[0214] 3. Surfers that are mentioned in the existing posts' content.

[0215] 4. Surfers having similar fields of interest based on the classifications made by the Knowledge Deduction Service.

[0216] 5. Posts that have a strong contextual matching, for example:

[0217] a. Same classification made by the Knowledge Deduction Service.

[0218] b. Same time in the time range of the posts in the case file.

[0219] c. Posts having a words-matching up to a certain threshold, etc.

[0220] It will be appreciated that the term "post" may be interpreted as any content distribution such as, publications, chats, content written by surfers, a product for sale, etc.

[0221] It will be appreciated by persons skilled in the art that the present invention is not limited to what has been particularly shown and described hereinabove. Rather the scope of the present invention is defined by the appended claims and includes combinations and sub-combinations of the various features described hereinabove as well as varia-

tions and modifications thereof which would occur to persons skilled in the art upon reading the foregoing description.

[0222] It is the intent of the applicant(s) that all publications, patents and patent applications referred to in this specification are to be incorporated in their entirety by reference into the specification, as if each individual publication, patent or patent application was specifically and individually noted when referenced that it is to be incorporated herein by reference. In addition, citation or identification of any reference in this application shall not be construed as an admission that such reference is available as prior art to the present invention. To the extent that section headings are used, they should not be construed as necessarily limiting. In addition, any priority document(s) of this application is/are hereby incorporated herein by reference in its/their entirety.

What is claimed is:

1. A method of providing searchable database and prioritized search user interface for exploring dark web content and surfer activity, comprising:

obtaining data comprising at least one web page scanned and collected from the dark web using a repository;

extracting from the data and storing in a structural database at least one structural parameter of a content of the at least one web page;

analyzing data in the structural database to determine and store in a search engine at least one statistic of the content of the at least one web page and to determine by employing machine learning and store in a knowledge database at least one profile characterization of at least one of a behavioral pattern of a surfer engaged with the content of the at least one web page and an interaction of the surfer with at least one other surfer;

for each of one or more search results of a search query to the structural database by the search engine, calculating a score according to a set of defined criteria using data stored in the search engine and the knowledge database; determining prioritization of the one or more search results using the score calculated for each; and

providing over a communication network to at least one computing device an output of the one or more search results prioritized according to the prioritization.

2. The method of claim 1, wherein the set of defined criteria comprising at least one of: source scoring; recency; user reputation; record type scoring; search result relevance scoring; and content analysis scoring.

3. The method of claim 1, wherein the at least one statistic comprising at least one of: date in which most of comments were written, number of posts a surfer wrote for a specific search query, distribution of categories in a site, time line trending for a specific search query and top sites for a specific query.

4. The method of claim 1, wherein determination of the at least one profile characterization comprising at least one of: analyzing sentiment of comments on posts to calculate reputation evaluation; classifying posts into categories and summing posts in each category to determine fields of interest; monitoring a number of interactions between surfers and identifying groups having a number of interactions above a predetermined threshold; and analyzing activity times.

5. The method of claim 4, wherein analyzing activity times comprising: calculating a temporal data distribution within a

time frame; storing time frame which includes most data; and storing an average and a standard deviation of the temporal data distribution.

6. The method of claim 1, wherein the analyzing further comprising identifying usage of different aliases by the at least one surfer using identity matching.

7. The method of claim 6, wherein the identity matching comprising at least one of:

locating communication information used by more than one surfer;

looking for similar aliases excluding common names;

locating surfers with similar activity pattern using activity times analysis;

locating surfers with similar fields of interest;

locating surfers who are active for a certain period and continue being active in other places or by other aliases;

locating surfers who post a same content at a same time in different locations;

counting most frequent words used by a surfer; and

analyzing surfers' text.

8. The method of claim 1, further comprising sending prioritized alerts according to rules defined and stored in an alert rules database responsive to using the search query within monitoring scheduled with relation to at least one alert.

9. The method of claim 8, wherein the alert rules database comprising at least one rule selected from the group consisting of: define wake up intervals for scheduling of monitoring with relation to the at least one alert, enable search by a key word, enable search by an activity related to a certain surfer, enable search by an activity of a certain group, enable search by a change in trend of a certain key word and enable search by a new phrase or a word that appears more than a predetermined number of times.

10. The method of claim 1, further comprising providing case management interface configured to enable a user to create a case file in order to manage a research or an investigation.

11. The method of claim 10, further comprising providing recommendation on adding relevant surfers and/or posts to the case file.

12. The method of claim 11, wherein the recommendation being provided according to at least one of: building a connection map of existing surfers in the case file, analyzing connections in the connection map and recommending adding surfers that have a strong connection with existing surfers in the case file; identity matching based similar surfers; surfers that published posts collected in the case file; surfers that are mentioned in existing posts' content; surfers having similar fields of interest; and posts that have a strong contextual matching comprising at least one of same classification, same time in a time range of posts in the case file and posts having a words-matching up to a certain threshold.

13. The method of claim 1, further comprising providing analytical dashboard to enable view of data analysis comprising at least one of categories, number of posts by dates, search results, an option to create an alert from a search, a total number of search results, surfer details, surfer activity analysis, surfer number of posts by dates, surfer categories and surfer connection map.

14. The method of claim 1, wherein obtaining the data from the dark web comprising using at least one crawler configured to manage an Internet Protocol address thereof by at least one of hiding the Internet Protocol address and changing the Internet Protocol address, progress from one web page to another using extracted links found in each web page, classify web

pages extracted and control operation timing and pace of data collection.

15. The method of claim 14, wherein the at least one crawler being further configured to optimize scanning pace versus secrecy thereof.

16. The method of claim 1, wherein the at least one profile characterization comprising reputation evaluation performed using formulas (I) and (II), wherein:

$$R(p) = \begin{cases} \left\{ \frac{S(q)}{R_0} \mid q \text{ in } Q_p \right\} * \sqrt[3]{|Q_p|} & \text{when } Q_p \neq \emptyset \\ 0 & \text{when } Q_p = \emptyset' \end{cases}$$

$$R(u) = \begin{cases} \left\{ \frac{R(p)}{G} \mid p \text{ in } P_u \right\} * \sqrt[3]{|P_u|} & \text{when } P_u \neq \emptyset \\ 0 & \text{when } P_u = \emptyset' \end{cases}$$

and wherein:

u = user;

p = post;

Pu = user post list;

Qp = post comments list, not including a user comments on its own post;

R(p) = post reputation;

R(u) = user reputation;

S(q) = sentiment of comment q, where  $-1 \leq S(q) \leq 1$ ;

R0 = reputation of a post with no comments;

(G) = average value of G;

|G| = number of members in G.

17. A system for searchable database and prioritized search user interface for exploring dark web content and surfer activity, comprising:

a repository configured to be used in obtaining data comprising at least one web page scanned and collected from the dark web;

a structured data extractor for extracting from the data and storing in a structural database at least one structural parameter of a content of the at least one web page;

an analytic engine for analyzing data in the structural database to determine and store in a search engine at least one statistic of the content of the at least one web page;

a knowledge deduction service for analyzing data in the structural database using machine learning to determine and store in a knowledge database at least one profile characterization of at least one of a behavioral pattern of a surfer engaged with the content of the at least one web page and an interaction of the surfer with at least one other surfer;

an alert service for:

calculating, for each of one or more search results of a search query to the structural database by the search engine, a score according to a set of defined criteria using data stored in the search engine and the knowledge database;

determining prioritization of the one or more search results using the score calculated for each; and providing over a communication network to at least one computing device an output of the one or more search results prioritized according to the prioritization.

18. The system of claim 17, further comprising at least one crawler and a hidden service locator connected with the dark web and the at least one crawler, the hidden service locator being configured to find at least one hidden uniform resource locator in the dark web; wherein the at least one crawler being configured to scan and collect from the dark web at least a



portion of said data using the at least one hidden uniform resource locator.

**19.** The system of claim **18**, further comprising a configuration database for storing respective of the at least one crawler a configuration comprising at least one of: at least one initial uniform resource location for the at least one crawler to start from, at least one username and at least one password of the at least one crawler, and timing setting of the at least one crawler.

**20.** The system of claim **17**, wherein the structured data extractor comprising:

- a wrapper generator;
- a wrapper database; and
- an extractor;

wherein the wrapper generator being configured to analyze each of the at least one web page, find patterns, create a wrapper and save the wrapper in the wrapper database; wherein the extractor being configured to receive a respective one of the at least one web page and a wrapper from the wrapper database and to extract data from the respective web page according the wrapper.

\* \* \* \* \*