(12) STANDARD PATENT
(19) AUSTRALIAN PATENT OFFICE

(11) Application No. **AU 2012217565 B2**

(54) Title
System and methods for identifying compromised personally identifiable information on the internet

(51) International Patent Classification(s)
G06F 11/00 (2006.01)

(21) Application No:	**2012217565**	(22) Date of Filing:	**2012.02.16**

(87) WIPO No: **WO12/112781**

(30) Priority Data

(31) Number	(32) Date	(33) Country
61/444,433	**2011.02.18**	**US**

(43) Publication Date:	**2012.08.23**
(44) Accepted Journal Date:	**2017.05.25**

(71) Applicant(s)
CSIdentity Corporation

(72) Inventor(s)
Gottschalk Jr., Harold E.;Caldwell, Michael;Carleton, Joel

(74) Agent / Attorney
Smoorenburg Patent and Trade Mark Attorneys, PO Box 515, RINGWOOD, VIC, 3134, AU

(56) Related Art
WO 2009/062111
US 2009/0328173
US 2010/0024037

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) **World Intellectual Property Organization**
International Bureau

(43) **International Publication Date**
23 August 2012 (23.08.2012)

WIPO | PCT

(10) **International Publication Number**
WO 2012/112781 A1

(51) **International Patent Classification**:
G06F 11/00 (2006.01)

(21) **International Application Number**:
PCT/US2012/025456

(22) **International Filing Date**:
16 February 2012 (16.02.2012)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
61/444,433 18 February 2011 (18.02.2011) US

(71) **Applicant** *(for all designated States except US)*: **CSIDENTITY CORPORATION** [US/US]; 7500 Rialto Blvd., Suite 260, Austin, TX 78735 (US).

(72) **Inventors; and**

(75) **Inventors/Applicants** *(for US only)*: **GOTTSCHALK, Harold, E., Jr.** [US/US]; 1870 Cordell Court, El Cajon, CA 92020 (US). **CALDWELL, Michael** [US/US]; 2644 Bristlewood Drive, Ramona, CA 92065 (US). **CARLETON, Joel** [US/US]; 14035 Barrymore Street, San Diego, CA 92129 (US).

(74) **Agents**: **ROBINSON, Ross, T.** et al.; Winstead PC, P.O. Box 131851, Dallas, TX 75313 (US).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: SYSTEM AND METHODS FOR IDENTIFYING COMPROMISED PERSONALLY IDENTIFIABLE INFORMATION ON THE INTERNET

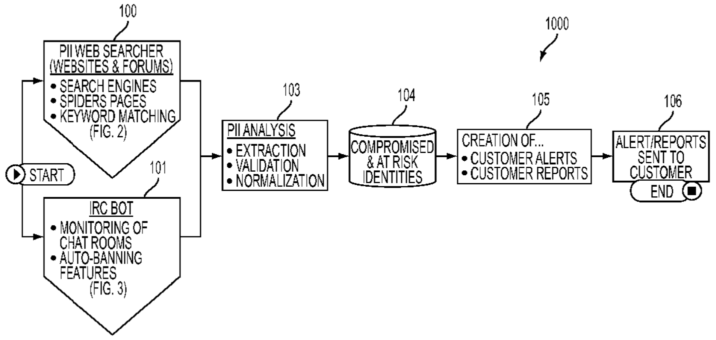

FIG. 1

(57) **Abstract**: In one embodiment, a method includes generating, by a computer system, a search-engine query from stored identity-theft nomenclature. The method also includes querying, by the computer system, at least one search engine via the search-engine query. Further, the method includes crawling, by the computer system, at least one computer-network resource identified via the querying. In addition, the method includes collecting, by the computer system, identity-theft information from the at least one computer-network resource. Additionally, the method includes processing, by the computer system, the identity-theft information for compromised personally-identifying information (PII).

SYSTEM AND METHODS FOR IDENTIFYING COMPROMISED PERSONALLY IDENTIFIABLE INFORMATION ON THE INTERNET

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority from, and incorporates by reference the entire disclosure of, U.S. Provisional Application No. 61/444,433 filed on February 18, 2011.

BACKGROUND

Technical Field

[0002] The present invention relates generally to the field of identity theft and more specifically, but not by way of limitation, to data mining of personally-identifying information found on the Internet.

History Of Related Art

[0003] Identity theft is a mounting concern in commercial transactions. This is particularly true in remote commercial transactions such as, for example, Internet-shopping transactions, that involve little or no direct personal contact between a consumer and a goods or services provider (GSP). It is commonplace for personally-identifying information (PII) to be compromised and utilized for identity theft such as, for example, in a remote commercial transaction. PII, as used herein, refers to information that can be used to uniquely identify, contact, or locate an individual person or can be used with other sources to uniquely identify, contact, or locate an individual person. PII may include, but is not limited to, social security numbers (SSN), bank or credit card account numbers, passwords, birth dates, and addresses. PII that has been obtained by or made available to a third party without proper authorization is referred to herein as compromised PII.

[0004] PII can be compromised in a myriad of ways. For example, record keeping for entities such as, for example, healthcare, governmental, financial, and educational institutions, is increasingly and sometimes exclusively electronic. Electronic record keeping introduces new risks for which the entities are frequently ill-equipped to handle. For example, PII is often compromised via stolen hardware, inadequate security procedures, security breaches, or employee carelessness or misconduct.

[0005] Another way that PII is frequently compromised is via "phishing." Phishing is the process of attempting to acquire PII by masquerading as a trustworthy entity in an electronic communication. A common example of phishing is a fraudulent email that is made to appear as though it originates from a valid source such as, for example, a national bank. The fraudulent email may incorporate a uniform resource locator (URL) that re-directs its audience to a false website that appears to be a legitimate website for the valid source. In actuality, the false website may be a front for stealing PII as part of a spurious transaction. For example, the false website may request "confirmation" of PII such as, for example, a credit card number or a username and password. The PII may then be stored for later improper use such as, for example, identity theft in a remote commercial transaction.

[0006] At least 182,395 instances of phishing were recorded during 2009, as reported by antiphishing.org. This is a forty-two percent increase over a number recorded in 2008. More than 10,745 malicious domains were registered in 2009, which is an increase of fifty-two percent over 2008. Sometimes, a misleading link such as, for example, the URL for the false website described above, may actually originate from a legitimate website but cause traffic to be redirected to an illegitimate website. This type of scam is known as "pharming."

[0007] Legislation to curb efforts to compromise PII are largely ineffective. For example, phishing and pharming activities originate from areas around the globe and are thus often protected from prosecution by a particular jurisdiction. Additionally, once PII is compromised, distribution of the compromised PII may be difficult or impossible to prevent. Web sites and forums dedicated to exchanging compromised PII are increasing rapidly in number. Some of these web sites and forums exchange compromised PII though email or secure direct uploads and downloads.

[0008] Identity theft resulting from compromised PII is costly to victims and companies alike. The Identity Fraud Survey Report created by Javelin Strategy & Research reported that in 2009 victims averaged a personal cost of $373 and 21 hours of time to resolve identity-theft issues. The annual cost of identity theft currently exceeds $200 billion worldwide. In addition, as a result of new legislation and litigation resulting from compromised PII, companies stand to suffer from lower profit margins, damaged credibility due to negative customer experiences, and eroded brand value. Identity theft also looms as a threat to the advancement of promising consumer-driven, self-service, and cost-savings technologies.

SUMMARY OF THE INVENTION

[0009] In one embodiment, a method includes generating, by a computer system, a search-engine query from stored identity-theft nomenclature. The method also includes querying, by the computer system, at least one search engine via the search-engine query. Further, the method includes crawling, by the computer system, at least one computer-network resource identified via the querying. In addition, the method includes collecting, by the computer system, identity-theft information from the at least one computer-network resource. Additionally, the method includes processing, by the computer system, the identity-theft information for compromised personally-identifying information (PII).

[00010] In one embodiment, a computer-program product includes a computer-usable medium having computer-readable program code embodied therein. The computer-readable program code is adapted to be executed to implement a method. The method includes generating, by a computer system, a search-engine query from stored identity-theft nomenclature. The method also includes querying, by the computer system, at least one search engine via the search-engine query. Further, the method includes crawling, by the computer system, at least one computer-network resource identified via the querying. In addition, the method includes collecting, by the computer system, identity-theft information from the at least one computer-network resource. Additionally, the method includes processing, by the computer system, the identity-theft information for compromised personally-identifying information (PII).

BRIEF DESCRIPTION OF THE DRAWINGS

[00011] A more complete understanding of the method and apparatus of the present invention may be obtained by reference to the following Detailed Description when taken in conjunction with the accompanying Drawings wherein:

[00012] FIG. 1 illustrates a process of identifying compromised PII on the Internet;

[00013] FIG. 2 illustrates a process of data mining for compromised PII using a PII Web Searcher;

[00014] FIG. 3 illustrates a process of data mining for compromised PII using an Internet Relay Chat Robot (IRC Bot);

[00015] FIG. 4 illustrates a process of chat room, nomenclature and website discovery;

[00016] FIG. 4A illustrates a process of nomenclature and website discovery;

[00017] FIG. 4B illustrates a process of chat-room discovery;

[00018] FIG. 5 illustrates a system that may be utilized to facilitate acquisition and utilization of identity-theft information; and

[00019] FIG. 6 illustrates an embodiment of a computer system on which various embodiments of the invention may be implemented.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS OF THE INVENTION

[00020] Although various embodiments of the method and apparatus of the present invention have been illustrated in the accompanying Drawings and described in the foregoing Detailed Description, it will be understood that the invention is not limited to the embodiments disclosed, but is capable of numerous rearrangements, modifications and substitutions without departing from the spirit of the invention as set forth herein.

[00021] FIG. 1 depicts an illustrative flow 1000 for identifying, analyzing, and reporting compromised PII on a computer network such as, for example, the Internet. In a typical embodiment, the flow 1000 may be initiated by one or both of a PII Web Searcher (PWS) 100 and an Internet Relay Chat Robot (IRC bot) 101. One of ordinary skill in the art will appreciate that the PWS 100 and the IRC bot 101 are illustrative in nature and that, in various embodiments, the flow 1000 may be initiated via other types of components that are operable to collect identity-theft information.

[00022] As used herein, *identity theft* generally involves a use of PII that is not authorized by an owner of the PII. Identity theft may include, for example, an unauthorized change to PII or an unauthorized use of PII to access resources or to obtain credit or other benefits. *Identity-theft information*, as used herein, includes any information that may be used to facilitate discovery or prevention of identity theft. Identity-theft information may include, for example, compromised PII and information related to where or how compromised PII may be found. *Identity-theft nomenclature*, as used herein, refers to words, phrases, nicknames, numbers, and the like that are determined to be suggestive of identity-theft information or identity theft. In various embodiments, identity-theft may include nomenclature for multiple languages (*e.g.*, English and non-English words).

[00023] In various embodiments, the flow 1000 may be initiated via the PWS 100. The PWS 100 may utilize, for example, search engines, web spiders, and keyword-matching features. In a typical embodiment, the search engines and the web spiders may be utilized to collect identity-theft information such as, for example, potential sources of compromised PII. The potential sources of compromised PII may include, for example, websites and forums that facilitate exchange of compromised PII (*e.g.*, by identity thieves). Further, keyword-matching features may be leveraged to analyze the potential sources of identity-theft information using, for example, identity-theft nomenclature. Additionally, the PWS 100 is generally operable to identify and collect other identity-theft information such as, for example, compromised PII, uniform resource locators (URLs), and references to IRC chat rooms (*i.e.*, channels). An illustrative embodiment of the PWS 100 will be described with respect to FIG. 2.

[00024] In various embodiments, the flow 1000 may be initiated via the IRC bot 101. Oftentimes, compromised PII is exchanged via chat rooms (*e.g.*, between identity thieves on IRC channels). In a typical embodiment, the IRC bot 101 is operable to crawl the Internet in search of chat rooms (*e.g.*, IRC channels) that are frequented by identity thieves. In a typical embodiment, the IRC bot 101 is operable to monitor such chat rooms for identity-theft nomenclature. Furthermore, the IRC bot 101 is typically operable to identify and collect compromised PII, URLs, references to other IRC chat rooms, and other identity-theft information from such chat rooms. Illustrative embodiments of the IRC bot 101 will be described with respect to FIGS. 3, 4, 4A, and 4B.

[00025] Oftentimes, if a particular user in a chat room is inactive for a certain period of time, the particular user may be timed out either automatically or by an administrator. In a typical embodiment, the IRC bot 101 may invoke auto-banning features that are operable to maintain an active status and thereby prevent time-out. The auto-banning features may involve simulating a human chat. For example, the auto-banning features may initiate a chat via a generic greeting, reproduce a single word from a monitored conversation, and the like. In a typical embodiment, the simulation of human chat may additionally cause an identity thief to reveal additional identity-theft information such as, for example, compromised PII or a URL to a potential source for compromised PII.

[00026] In various embodiments, the IRC bot 101 and the PWS 100 may operate collaboratively in the flow 1000. For example, the IRC bot 101 may provide identity-theft nomenclature such as email addresses, nicknames, and other information that may be used by an identity thief. The IRC bot 101 may further provide, for example, URLs to potential sources of compromised PII. In a typical embodiment, the PWS 100 may crawl the URLs provided by the IRC bot 101 and scan for identity-theft information. The PWS 100 may also search and crawl the Internet using the identity-theft nomenclature provided by the IRC bot 101. In a similar manner, the PWS 100 may discover and send identity-theft information such as, for example, chat rooms, to the IRC bot 101. In a typical embodiment, the IRC bot 101 may monitor the chat rooms provided by the PWS 100.

[00027] After identity-theft information is collected by the IRC bot 101 and the PWS 100, the collected identity-theft information may be processed at step 103. In a typical embodiment, the processing of the collected identity-theft information may include an extraction process, a validation process, and a normalization process. In various embodiments, the PWS 100 and the IRC bot 101 may yield extensive amounts of identity-theft information that includes, for example, webpage segments, IRC logs, text files, and the like. In a typical embodiment, the extraction process and the validation process operate to intelligently reduce an amount of the collected identity-theft information that is stored and utilized in subsequent steps of the flow 1000. In a typical embodiment, the normalization process ensures that the identity-theft information is stored efficiently and effectively.

[00028] In a typical embodiment, as part of the extraction process, the collected identity-theft information may be processed for compromised PII by one or more parsers that recognize common formats for PII. For example, a parser may identify token-separated data (*e.g.*, tab-delimited data). Similarly, a parser may determine a column type for columns lacking a column header, for example, by analyzing data that is present in particular columns (*e.g.*, recognizing a list of text strings as email addresses). Furthermore, a parser may identify multi-line labeled data such as, for example, "first name: John," and various other labels that may be associated with compromised PII (*e.g.*, recognizing "ccn," "cc" or "credit card" as possible labels for credit-card information). Additionally, by way of further example, a parser may identify identity-theft information taken from encodings that may be present on cards such as, for example, credit cards, driver's licenses, and the like. The encodings may include, for example, track 1 and track 2 magnetic-stripe data.

[00029] Additionally, as part of the extraction process, rules may be enforced that require groups of fields to be present in particular compromised PII before allowing the particular compromised PII to be recorded. In a typical embodiment, the requirement that groups of fields be present has the benefit of reducing "false positives" within compromised PII. False positives may be considered elements of compromised PII that are not deemed to be sufficiently private or sufficiently important to merit recordation. In a typical embodiment, false positives may be removed from the collected identity-theft information. For example, an email address that is not accompanied by a password may be considered a false positive and not recorded. In a typical embodiment, a rule may be established that requires, for example, a username or email address to be accompanied by a password in order to be recorded.

[00030] In a typical embodiment, the validation process involves analyzing a source of the collected identity-theft information such as, for example, compromised PII, and determining if any elements of the compromised PII are false positives. For example, in a typical embodiment, genealogy websites, phone/address lookup websites, and website log files are common sources of false positives. Compromised PII that is mined from such websites, in a typical embodiment, may be considered false positives and removed from the collected identity-theft information. Conversely, compromised PII mined, for example, from known hacker websites and websites replete with identity-theft nomenclature, in a typical embodiment, may be protected from identification as false positives.

[00031] In a typical embodiment, the normalization process ensures that the collected identity-theft information such as, for example, compromised PII, is stored according to a standardized format. For example, standardized data structures and attributes may be established for names, credit-card numbers, and the like. In a typical embodiment, the normalization process facilitates matching, for example, elements of compromised PII to particular individuals to whom the elements correspond. In that way, reports and alerts based on the compromised PII may be more efficiently and more accurately generated. In a typical embodiment, after the extraction process, the validation process, and the normalization process, the collected identity-theft information is recorded in a database at step 104.

[00032] At step 105, in a typical embodiment, alerts and reports may be delivered based on, for example, compromised PII that is stored in the database at step 104. In some embodiments, the recordation of any elements of compromised PII at step 104 merits delivery of an alert to an individual to whom the elements correspond. In other embodiments, an individual may only be delivered an alert if, for example, certain elements or combinations of elements are discovered and recorded (*e.g.*, credit-card information or social-security-number). In a typical embodiment, a particular individual may be able to pre-specify an alert-delivery method (*e.g.*, email, telephone, etc.). After step 105, the flow 1000 ends.

[00033] FIG. 2 illustrates a flow 2000 for mining compromised PII via a PWS 200. In a typical embodiment, the PWS 200 is similar to the PWS 100 of FIG. 1. The PWS 200 typically accesses a database 203 that includes identity-theft nomenclature and identity-theft websites. Identity-theft websites are websites that have been identified via, for example, identity-theft nomenclature, to be possible sources of compromised PII. The database 203 is typically populated with identity-theft websites and identity-theft nomenclature via a discovery process 204. Illustrative embodiments of the discovery process 204 will be described in further detail with respect to FIG. 4A.

[00034] In a typical embodiment, the PWS 200 receives identity-theft nomenclature and identity-theft websites as input from the database 203. The PWS 200 typically queries search engines 201 via keywords from the identity-theft nomenclature. Additionally, the PWS 200 typically crawls websites 202 and scans the websites 202 for the identity-theft nomenclature. In a typical embodiment, the websites 202 include the identity-theft websites received as input from the database 203 and websites identified via queries to the search engines 201. At step 206, compromised PII collected by the PWS 200 may be processed at a processing step 206 in a manner similar to that described with respect to step 103 of FIG. 1.

[00035] As new websites and identity-theft nomenclature are added to the database 203 via, for example, the discovery process 204, the database 203 may be optimized via a performance-analysis process 205. In the performance-analysis 205, the identity-theft nomenclature is typically ranked according to a relative significance of compromised PII that is gleaned thereby. In a typical embodiment, the database 203 maintains, for each element of the identity-theft nomenclature, historical information related to compromised PII obtained via that element. In a typical embodiment, each element of the identity-theft nomenclature may be ranked, for example, according to an amount and/or a quality of the compromised PII obtained via that element.

[00036] The quality of the compromised PII may be determined, for example, by assigning weights based on a degree of sensitivity of particular elements of compromised PII. For example, in various embodiments, credit-card information and social security numbers may be assigned higher weights than, for example, website account information. In various

embodiments, the amount of compromised PII may be, for example, an overall amount of compromised PII historically obtained via particular identity-theft nomenclature. Further, in various embodiments, the amount of compromised PII may be, for example, an amount of PII obtained via particular identity-theft nomenclature in a defined period of time. For example, in some embodiments, it may be advantageous to consider an amount of compromised PII obtained via particular identity-theft nomenclature within the last thirty days.

[00037] In a typical embodiment, a score may be computed for each element of identity-theft nomenclature based on, for example, an amount and/or a quality of the compromised PII that is gleaned thereby. In a typical embodiment, a scoring formula for generating the score is configurable. For example, weighting factors may be assigned to the amount and/or the quality of the compromised PII. In that way, greater or less weight may be assigned to the amount and/or the quality of the compromised PII, as may be desired for particular applications. Once scores are generated for each element of the identity-theft nomenclature, the identity-theft nomenclature may be ranked based on the scores.

[00038] In a typical embodiment, the PWS 200 may query the search engines 201 via keywords from the ranked identity-theft nomenclature in order to yield, for example, URLs to additional websites. The additional websites may be stored in the database 203. In a typical embodiment, the PWS 200 may crawl and scan the additional websites in a manner similar to that described above with regard to the websites 202. Further, compromised PII collected by the PWS 200 may be processed at a processing step 206 in a manner similar to that described with respect to step 103 of FIG. 1. After the performance-analysis process 205, the flow 2000 ends.

[00039] FIG. 3 illustrates a flow 3000 for compiling databases of compromised PII via an IRC bot. The flow 3000 begins via a chat-room-discovery process 300. During the chat-room-discover process 300, a database 301 is populated. The database 301, in a typical embodiment, includes URLs, for example, to IRC networks and channels likely to relate to identity theft. An illustrative embodiment of the chat-room-discovery process 300 will be described in more detail with respect to FIG. 4B.

[00040] In a typical embodiment, an IRC bot 302 receives URLs for IRC networks 303 as input from the database 301. The IRC bot 302 is generally similar to the IRC bot 101 of FIG. 1. The IRC bot 302 typically scans the IRC networks 303 for identity-theft information such as, for example, compromised PII. In a typical embodiment, the IRC bot 302 invokes one or more auto-banning features 304 in order to prevent being timed out on a particular IRC network due to inactivity. For example, the IRC bot 304 may simulate human interaction by interjecting text. In a typical embodiment, the IRC bot 304 is further operable to change Internet Protocol (IP) addresses in order explore IRC networks and chat rooms with efficiency.

[00041] Any compromised PII that is found by the IRC bot 302 is typically logged into an IRC log database 305. After being logged, in a typical embodiment, the compromised PII is processed at a processing step 306 in a manner similar to that described with respect to step 103 of FIG. 1. After the processing step 306, the flow 3000 ends.

[00042] FIG. 4 depicts an illustrative flow 4000 for chat room and website discovery. In particular, the flow 4000 illustrates interactions between an IRC bot 400, a chat-room-discovery process 405, a nomenclature-and-website discovery process 404, a dialog-extraction process 402, an IRC log database 401, and an IRC dialog database 403. The IRC bot 400 is generally operable to scan chat rooms on IRC networks for compromised PII. In a typical embodiment, the IRC bot 400 is similar to the IRC bot 101 of FIG. 1 and the IRC bot 302 of FIG. 3.

[00043] After the chat rooms are scanned by the IRC bot 400 as described with respect to FIGS. 1 and 3, identity-theft information such as, for example, compromised PII, is typically logged into the IRC log database 401 as an IRC log. In a typical embodiment, the dialog-extraction process 402 is applied to the IRC log. The dialog-extraction process 402 is typically similar to the extraction process described with respect to step 103 of FIG. 1. In a typical embodiment, compromised PII that is extracted as part of the dialog-extraction process is stored in the IRC dialog database 403. In a typical embodiment, automated spam postings can be distinguished and separated from other dialog.

[00044] In a typical embodiment, the IRC log stored in the IRC log database 401 and the extracted compromised PII stored in the IRC dialog database 403 may be provided as inputs to the nomenclature-and-website discover process 404. In a typical embodiment, the nomenclature-and-website discover process 404 discovers new websites and identity-theft nomenclature that may be utilized, for example, by the IRC bot 400, to acquire additional identity-theft information. An illustrative embodiment of the nomenclature-and-website discovery process 404 will be described in more detail with respect to FIG. 4A.

[00045] In a typical embodiment, the IRC log stored in the IRC log database 401 may be provided as input to the chat-room-discovery process 405. Although not illustrated, in various embodiments, the extracted compromised PII stored in the IRC dialog database 403 may also be provided as input to the chat-room-discovery process 405. In a typical embodiment, the chat-room-discovery process 405 analyzes the IRC log in order to identify, for example, references to new chat rooms on IRC networks that may be sources of compromised PII. An illustrative embodiment of the chat-room-discovery process 405 will be described with respect to FIG. 4B.

[00046] FIG. 4A is an illustrative flow 4000A for nomenclature and website discovery. The flow 4000A typically begins with an IRC bot 400A. In a typical embodiment, the IRC bot 400A is similar to the IRC bot 400 of FIG. 4, the IRC bot 300 of FIG. 3, and the IRC bot 101 of FIG. 1. At a discovery step 401A, an IRC log generated by the IRC bot 400A may be analyzed for new identity-theft nomenclature and new websites. The IRC log may be, for example, an IRC log from the IRC log database 401 of FIG. 4. The new identity-theft nomenclature may include, for example, nicknames and email addresses used by participants (*e.g.*, identity thieves) in chat rooms. By way of further example, the new websites may include URLs to websites that are mentioned in chat rooms. In various embodiments, the new identity-theft nomenclature may be utilized by a PWS such as, for example, the PWS 200 of FIG. 2, to search for additional compromised PII as described with respect to FIG. 2.

[00047] After the discovery step 401A, an analysis step 402A may occur. In a typical embodiment, the analysis step 402A includes ranking a relative significance of identity-theft websites and forums that are stored, for example, in a database 403A. The identity-theft websites and forums include, for example, the new websites and forums identified at the discovery step 401A. The identity-theft websites and forums may be ranked in a manner similar to that described with respect to the ranking of identity-theft nomenclature in the performance-analysis process 205 of FIG. 2. In a typical embodiment, the analysis step 402A results in storage of the rankings and the new websites in the database 403A. Subsequently, the flow 4000A ends.

[00048] FIG. 4B illustrates a flow 4000B for chat-room discovery. In a typical embodiment, the flow 4000B may begin via an IRC bot 400B. In a typical embodiment, the IRC bot 400B is similar to the IRC bot 400A of FIG. 4A, the IRC bot 400 of FIG. 4, the IRC bot 300 of FIG. 3, and the IRC bot 101 of FIG. 1. As described with respect to FIG. 4, the IRC bot 400B may yield IRC logs from monitoring of chat rooms. Additionally, as described with respect to the discovery process 401A of FIG. 4A, in various embodiments, the IRC bot 400B may yield identity-theft nomenclature and identity-theft websites after engaging in a discovery process. The identity-theft nomenclature and the identity-theft websites may be stored, for example, in a nomenclature database 403B and an IRC-network database 404B.

[00049] In a typical embodiment, the IRC logs, the identity-theft nomenclature from the nomenclature database 403B and the chat rooms from the chat-room database 404B may serve as inputs to an analysis step 401B. At the analysis step 401B, the flow 4000B is typically operable to analyze the IRC logs to discover new chat rooms. For example, for a given IRC log, the flow 4000B may analyze a frequency of identity-theft nomenclature. In addition, by way of further example, the flow 4000B may determine how often particular chat rooms are referenced in a given IRC log. In various embodiments, if references to a particular chat room exceed a configurable threshold, the particular chat room may be recorded in a database 405B at step 402B. In some embodiments, the predetermined threshold for overall references may vary based on, for example, a frequency of identity-theft nomenclature in the given IRC log. For example, if the given IRC log has a high frequency of identity-theft nomenclature relative to a configurable value, a single reference may be sufficient for recordation in the database 405B.

[00050] In various embodiments, the analysis step 401B may further involve monitoring particular chat rooms from the chat-room database 404B. For example, as described with respect to the analysis step 402A of FIG. 4A, chat rooms in the chat-room database 404B may be ranked. Therefore, in various embodiments, high-ranking chat rooms may be monitored for references to other chat rooms. In a typical embodiment, new chat rooms discovered via the analysis step 401B are stored in the database 405B at step 402B. Subsequently, the flow 4000B ends.

[00051] FIG. 5 illustrates a system 500 that may be utilized to facilitate acquisition and utilization of identity-theft information. The system 500 includes a server computer 502, a database 504, and a computer network 506. In a typical embodiment, the server computer 502 may have resident and operating thereon a PWS such as, for example, the PWS 200 of FIG. 2. In a typical embodiment, the server computer may have resident and operating thereon an IRC bot such as, for example, the IRC bot 400B of FIG. 4B, the IRC bot 400A of FIG. 4A, the IRC bot 400 of FIG. 4, the IRC bot 300 of FIG. 3, and the IRC bot 101 of FIG. 1. In various embodiments, the server computer 502 may facilitate execution, for example, of the flow 1000 of FIG. 1, the flow 2000 of FIG. 2, the flow 3000 of FIG. 3, and/or the flow 4000 of FIG. 4. In that way, the server computer 502 may be operable to acquire identity-theft information such as, for example, compromised PII, via the computer network 506. The computer network 506 may be, for example, the Internet. The identity-theft information may be stored, for example, in the database 504.

[00052] One of ordinary skill in the art will appreciate that the server computer 502 may, in various embodiments, represent a plurality of server computers. For example, the PWS and the IRC bot may, in various embodiments, be resident and operating on distinct physical or virtual server computers. Likewise, in various embodiments, the PWS and the IRC bot may be resident and operating on one physical or virtual server computer. Furthermore, one of ordinary skill in the art will appreciate that the database 504 may, in various embodiments, represent either a single database or a plurality of databases.

[00053] FIG. 6 illustrates an embodiment of a computer system 600 on which various embodiments of the invention may be implemented such as, for example, the PWS 200 of FIG. 2, the IRC bot 400B of FIG. 4B, the IRC bot 400A of FIG. 4A, the IRC bot 400 of FIG. 4, the IRC bot 300 of FIG. 3, and the IRC bot 101 of FIG. 1. The computer system 600 may be, for example, similar to the server computer 502 of FIG. 5. The computer system 600 may be a physical system, virtual system, or a combination of both physical and virtual systems. In the implementation, a computer system 600 may include a bus 618 or other communication mechanism for communicating information and a processor 602 coupled to the bus 618 for processing information. The computer system 600 also includes a main memory 604, such as random-access memory (RAM) or other dynamic storage device, coupled to the bus 618 for storing computer readable instructions by the processor 602.

[00054] The main memory 604 also may be used for storing temporary variables or other intermediate information during execution of the instructions to be executed by the processor 602. The computer system 600 further includes a read-only memory (ROM) 606 or other static storage device coupled to the bus 618 for storing static information and instructions for the processor 602. A computer-readable storage device 608, such as a magnetic disk or optical disk, is coupled to the bus 618 for storing information and instructions for the processor 602. The computer system 600 may be coupled via the bus 618 to a display 610, such as a liquid crystal display (LCD) or a cathode ray tube (CRT), for displaying information to a user. An input device 612, including, for example, alphanumeric and other keys, is coupled to the bus 618 for communicating information and command selections to the processor 602. Another type of user input device is a cursor control 614, such as a mouse, a trackball, or cursor direction keys for communicating direct information and command selections to the processor 602 and for controlling cursor movement on the display 610. The cursor control 614 typically has two degrees of freedom in two axes, a first axis (e.g., x) and a second axis (e.g., y), that allow the device to specify positions in a plane.

[00055] The term "computer readable instructions" as used above refers to any instructions that may be performed by the processor 602 and/or other component of the computer system 600. Similarly, the term "computer readable medium" refers to any storage medium that may be used to store the computer readable instructions. Such a medium may take many forms, including, but not limited to, non volatile media, volatile media, and transmission media. Non-volatile media include, for example, optical or magnetic disks, such as the storage device 608. Volatile media includes dynamic memory, such as the main memory 604. Transmission media includes coaxial cables, copper wire, and fiber optics, including wires of the bus 618.

Transmission media can also take the form of acoustic or light waves, such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD ROM, DVD, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, a PROM, an EPROM, a FLASH EPROM, any other memory chip or cartridge, a carrier wave, or any other medium from which a computer can read.

[00056] Various forms of the computer readable media may be involved in carrying one or more sequences of one or more instructions to the processor 602 for execution. For example, the instructions may initially be borne on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to the computer system 600 can receive the data on the telephone line and use an infrared transmitter to convert the data to an infrared signal. An infrared detector coupled to the bus 618 can receive the data carried in the infrared signal and place the data on the bus 618. The bus 618 carries the data to the main memory 604, from which the processor 602 retrieves and executes the instructions. The instructions received by the main memory 604 may optionally be stored on the storage device 608 either before or after execution by the processor 602.

[00057] The computer system 600 may also include a communication interface 616 coupled to the bus 618. The communication interface 616 provides a two-way data communication coupling between the computer system 600 and a network. For example, the communication interface 616 may be an integrated services digital network (ISDN) card or a modem used to provide a data communication connection to a corresponding type of telephone line. As another example, the communication interface 616 may be a local area network (LAN) card used to provide a data communication connection to a compatible LAN. Wireless links may also be implemented. In any such implementation, the communication interface 616 sends and receives electrical, electromagnetic, optical, or other signals that carry digital data streams representing various types of information. The storage device 608 can further include instructions for carrying out various processes for image processing as described herein when executed by the processor 602. The storage device 608 can further include a database for storing data relative to same.

"Comprises/comprising" when used in this specification is taken to specify the presence of stated features, integers, steps or components but does not preclude the presence or addition of one or more other features, integers, steps, components or groups thereof."
The discussion throughout this specification comes about due to the realisation of the inventors and/or the identification of certain prior art problems by the inventors.

Any discussion of documents, devices, acts or knowledge in this specification is included to explain the context of the invention. It should not be taken as an admission that any of the material forms a part of the prior art base or the common general knowledge in the relevant art in Australia or elsewhere on or before the priority date of the disclosure and claims herein.

CLAIMS

What is claimed is:

1. A method comprising:

accessing, by a computer system, stored identity-theft nomenclature;

wherein the stored identity-theft nomenclature comprises a changing set of words and phrases determined to be suggestive of at least one of:

identity theft; and

exchange of identity-theft information;

generating, by a computer system, a search-engine query from stored identity-theft nomenclature;

querying, by the computer system, at least one search engine via the search-engine query;

identifying, by the computer system, at least one new computer-network resource responsive to the querying; crawling, by the computer system, the at least one computer-network resource;

collecting, by the computer system, identity-theft information from the at least one computer-network resource; and

processing, by the computer system, the identity-theft information for compromised personally-identifying information (PII); and

wherein the processing comprises:

analyzing the identity-theft information for new identity-theft nomenclature; and

storing any new identity-theft nomenclature with the stored identity-theft nomenclature.

2. The method of claim 1, wherein the collecting comprises scanning the at least one computer-network resource for at least a portion the stored identity-theft nomenclature.

3. The method of claim 1, wherein the processing comprises extracting compromised PII from the identity-theft information, the extracting comprising recognizing at least one PII format.

4. The method of claim 3, wherein the recognized PII format is selected from the group consisting of: token-separated data, one or more columns of data lacking column headers, multi-line labeled data, and magnetic-stripe data.

5. The method of claim 1, wherein the processing comprises validating the at least one computer-network resource, the validating comprising determining whether the at least one computer-network resource is a likely source of false positives for compromised PII.

6. The method of claim 1, wherein the processing comprises normalizing the identity-theft information, the normalizing comprising storing the identity-theft information according to a standardized format.

7. The method of claim 1, comprising creating and delivering at least one of an alert and a report in connection with the identity-theft information.

8. The method of claim 1, wherein the identity-theft information comprises information related to new sources of compromised PII.

9. The method of claim 1, wherein the stored identity-theft nomenclature comprises words that are determined to be suggestive of identity-theft information.

10. The method of claim 1, comprising ranking entries within the stored identity-theft nomenclature according to a relative significance of compromised PII that is gleaned thereby.

11. The method of claim 10, wherein the ranking comprises ranking the stored identity-theft nomenclature according to a quality of compromised PII that is gleaned thereby.

12. The method of claim 10, wherein the ranking comprises ranking the stored identity-theft nomenclature according to a quantity of compromised PII that is gleaned thereby.

13. The method of claim 10, wherein the generating comprises generating the search- engine query from highly-ranked entries from the stored identity-theft nomenclature.

14. The method of claim 1, wherein the at least one computer-network resource is a chat room.

15. The method of claim 14, wherein the collecting comprises distinguishing spam postings from other dialog.

16. The method of claim 14, wherein the collecting comprises logging chat dialog into a chat log database.

17. The method of claim 16, wherein the processing comprises discovering new chat rooms, the discovering comprising analyzing chat dialogs stored in the chat log database.

18. The method of claim 17, wherein the discovering comprises analyzing a frequency of the stored identity-theft nomenclature in the chat dialogs.

19. The method of claim 18, wherein the discovering comprises determining how often particular chat rooms are referenced in a given chat dialog from the chat log database.

20. The method of claim 19, wherein the discovering comprises, responsive to references to a given chat room exceeding a threshold, recording the given chat room for future crawling.

21. The method of claim 1, wherein the identity-theft information identifies at least one of a chat network and a chat room that is determined likely to relate to identity theft.

22. The method of claim 1, wherein the identity-theft information comprises a uniform resource locator (URL) to a website that is determined likely to relate to identity theft.

23. The method of claim 1, wherein the identity-theft nomenclature comprises non- English words.

24. A computer-program product comprising a non-transitory computer-usable medium having computer-readable program code embodied therein, the computer-readable program code

adapted to be executed to implement a method comprising:

accessing stored identity-theft nomenclature;

wherein the stored identity-theft nomenclature comprises a changing set of words and phrases determined to be suggestive of at least one of:

identity theft; and

exchange of identity-theft information; generating a search-engine query from stored identity-theft nomenclature;

querying at least one search engine via the search-engine query;

identifying at least one new computer-network resource responsive to the querying;

crawling the at least one computer-network resource identified via the querying;

collecting identity-theft information from the at least one computer-network resource;

processing the identity-theft information for compromised personally-identifying information (PII); and

wherein the processing comprises:

analyzing the identity-theft information for new identity-theft nomenclature; and

storing any new identity-theft nomenclature with the stored identity-theft nomenclature.

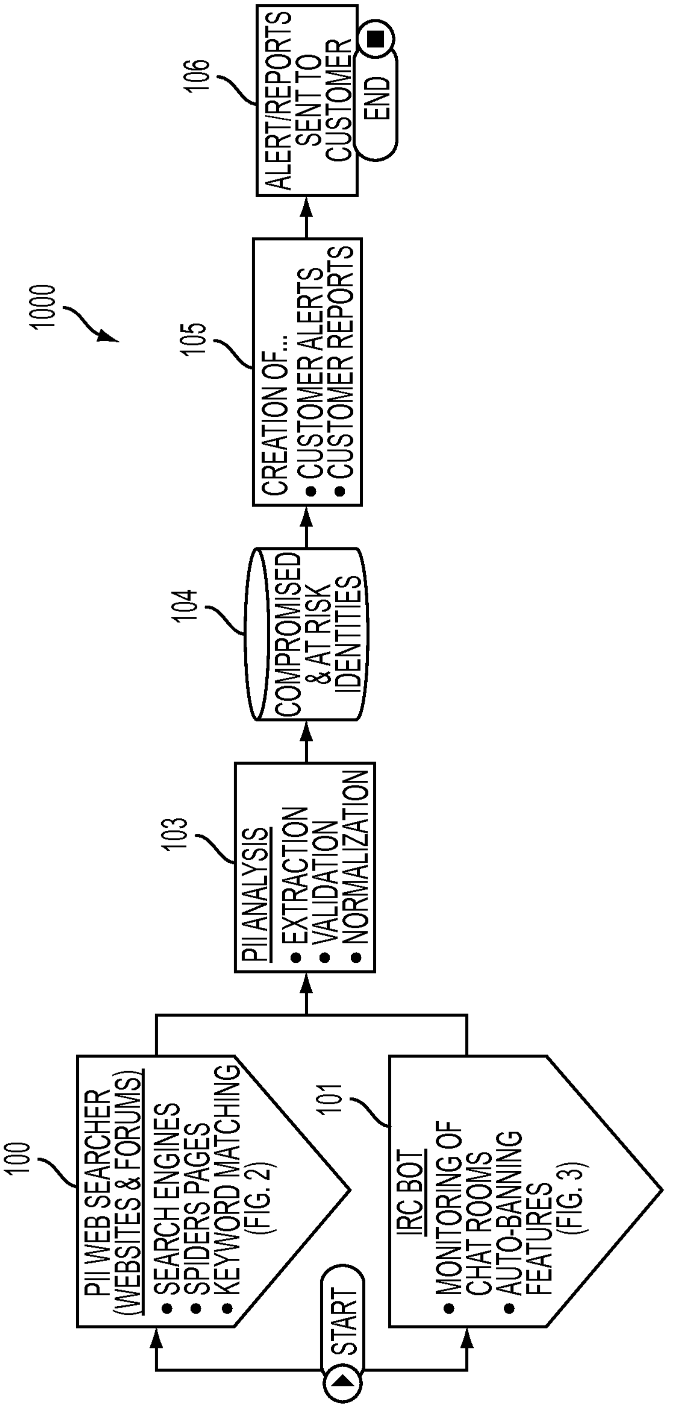
START
100
PII WEB SEARCHER
(WEBSITES & FORUMS)
• SEARCH ENGINES
• SPIDERS PAGES
• KEYWORD MATCHING
(FIG. 2)
101
IRC BOT
• MONITORING OF CHAT ROOMS
• AUTO-BANNING FEATURES
(FIG. 3)
103
PII ANALYSIS
• EXTRACTION
• VALIDATION
• NORMALIZATION
104
COMPROMISED & AT RISK IDENTITIES
1000
105
CREATION OF...
• CUSTOMER ALERTS
• CUSTOMER REPORTS
106
ALERT/REPORTS SENT TO CUSTOMER
END

FIG. 1

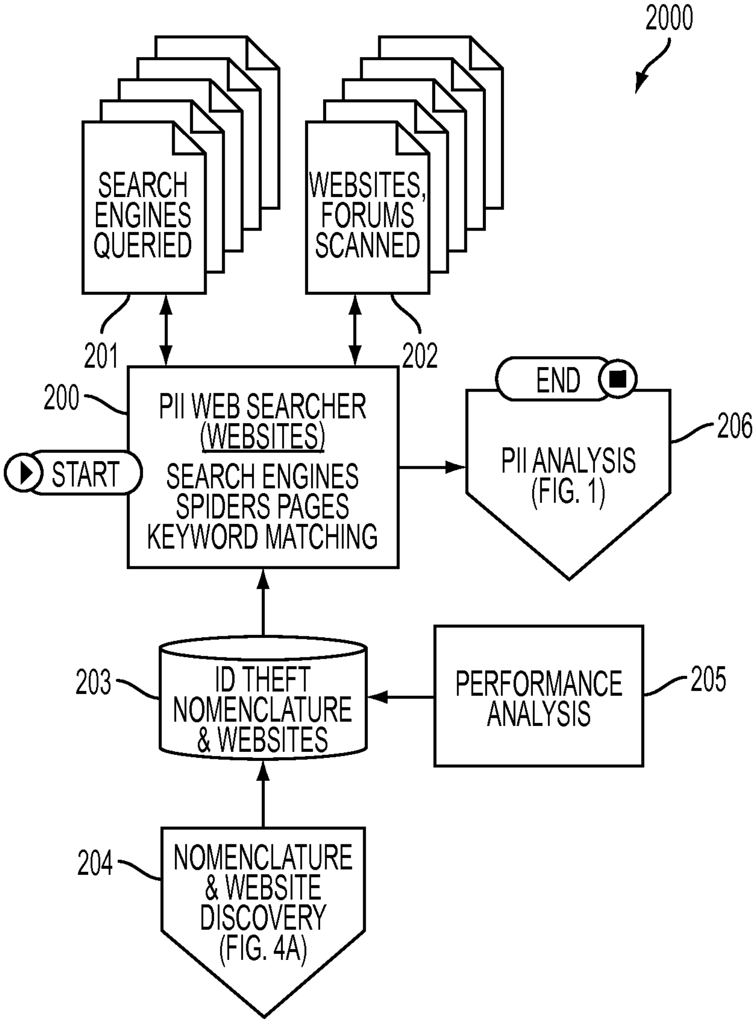
2000
SEARCH ENGINES QUERIED
WEBSITES, FORUMS SCANNED
201
202
200
PII WEB SEARCHER (WEBSITES)
SEARCH ENGINES SPIDERS PAGES KEYWORD MATCHING
START
END
PII ANALYSIS (FIG. 1)
206
203
ID THEFT NOMENCLATURE & WEBSITES
PERFORMANCE ANALYSIS
205
204
NOMENCLATURE & WEBSITE DISCOVERY (FIG. 4A)

FIG. 2

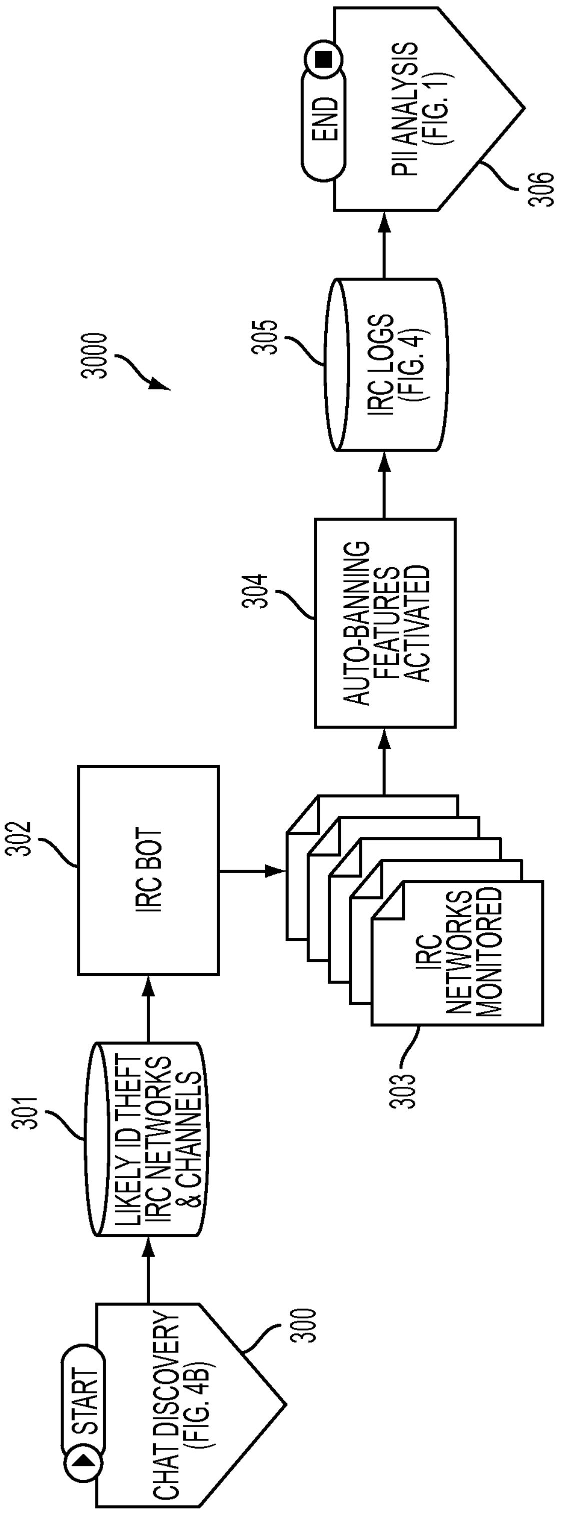
3000
START
CHAT DISCOVERY (FIG. 4B)
300
LIKELY ID THEFT IRC NETWORKS & CHANNELS
301
IRC BOT
302
IRC NETWORKS MONITORED
303
AUTO-BANNING FEATURES ACTIVATED
304
IRC LOGS (FIG. 4)
305
PII ANALYSIS (FIG. 1)
306
END

FIG. 3

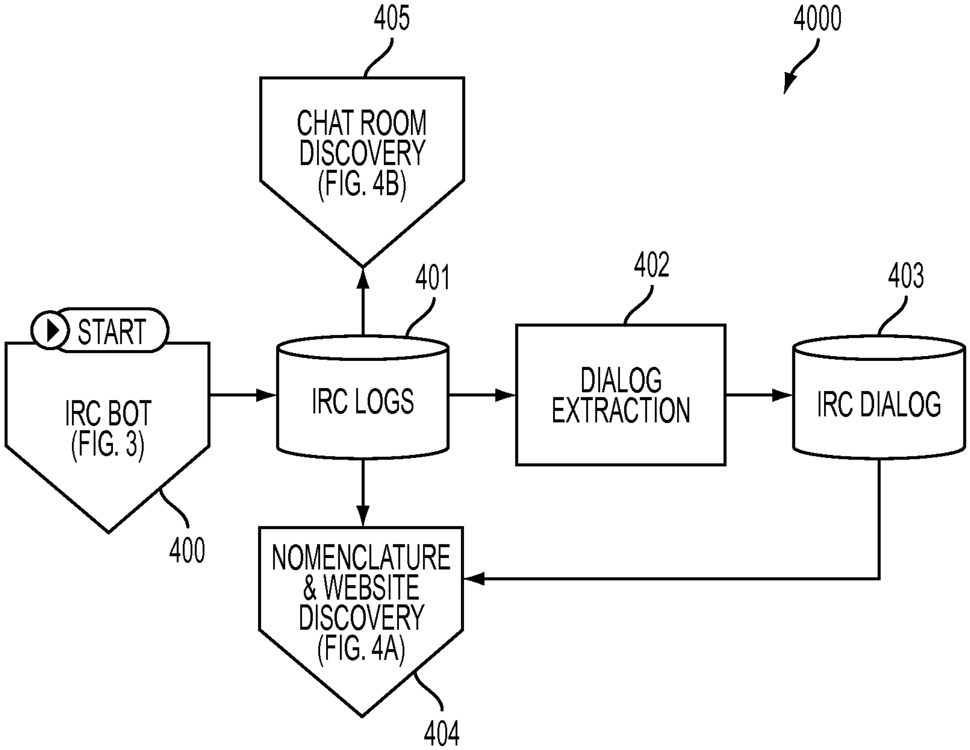
405
4000
CHAT ROOM DISCOVERY (FIG. 4B)
401
402
403
START
IRC BOT (FIG. 3)
IRC LOGS
DIALOG EXTRACTION
IRC DIALOG
400
NOMENCLATURE & WEBSITE DISCOVERY (FIG. 4A)
404

FIG. 4

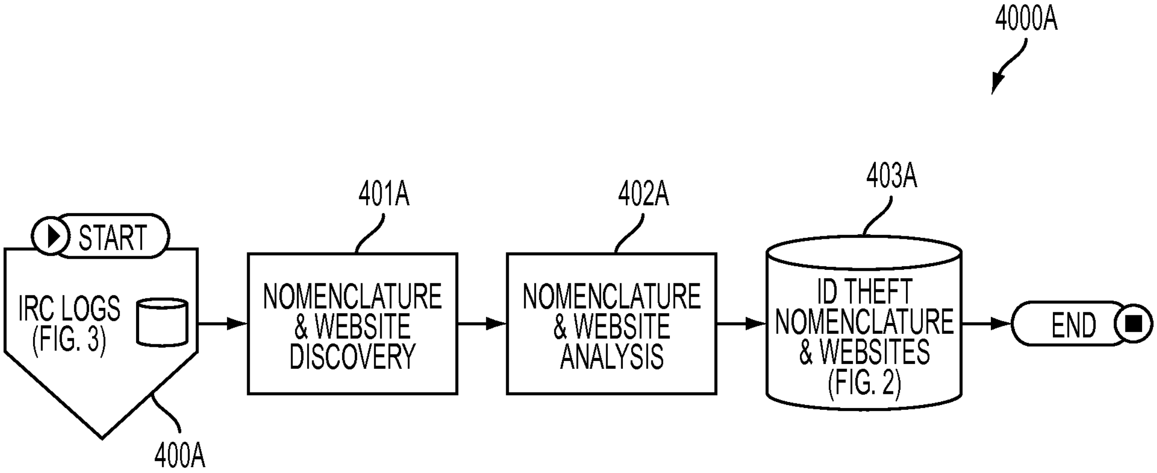
4000A
START
IRC LOGS (FIG. 3)
400A
401A
NOMENCLATURE & WEBSITE DISCOVERY
402A
NOMENCLATURE & WEBSITE ANALYSIS
403A
ID THEFT NOMENCLATURE & WEBSITES (FIG. 2)
END

FIG. 4A

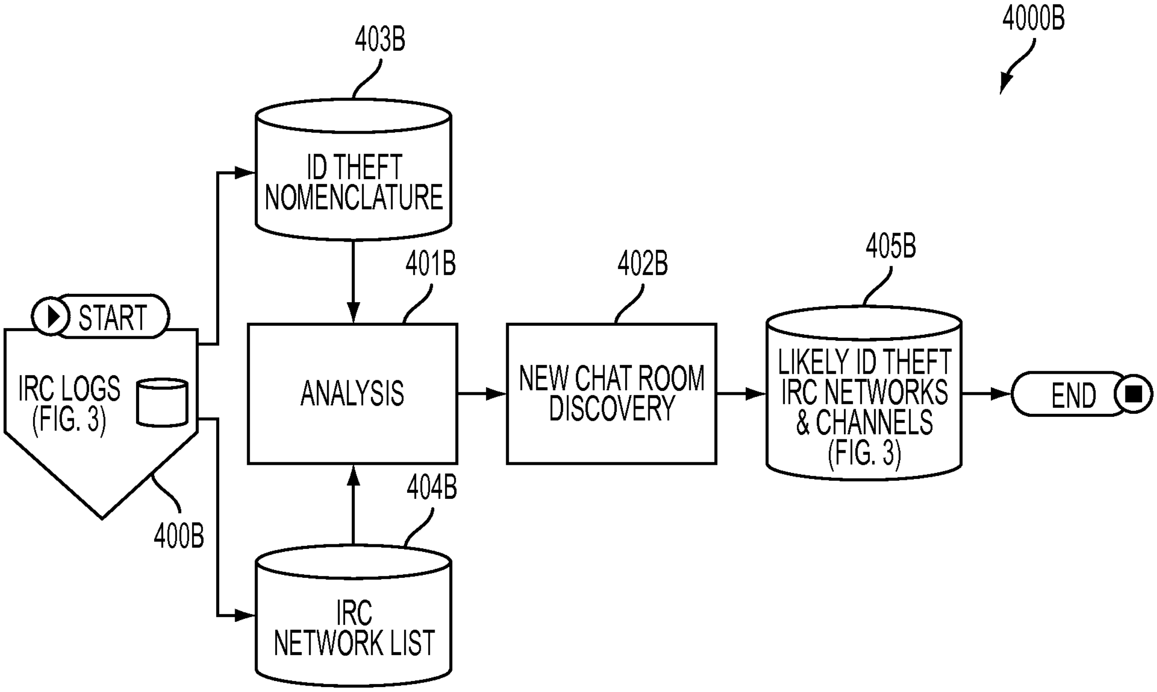
4000B
403B
ID THEFT NOMENCLATURE
START
IRC LOGS (FIG. 3)
400B
401B
ANALYSIS
404B
IRC NETWORK LIST
402B
NEW CHAT ROOM DISCOVERY
405B
LIKELY ID THEFT IRC NETWORKS & CHANNELS (FIG. 3)
END

FIG. 4B

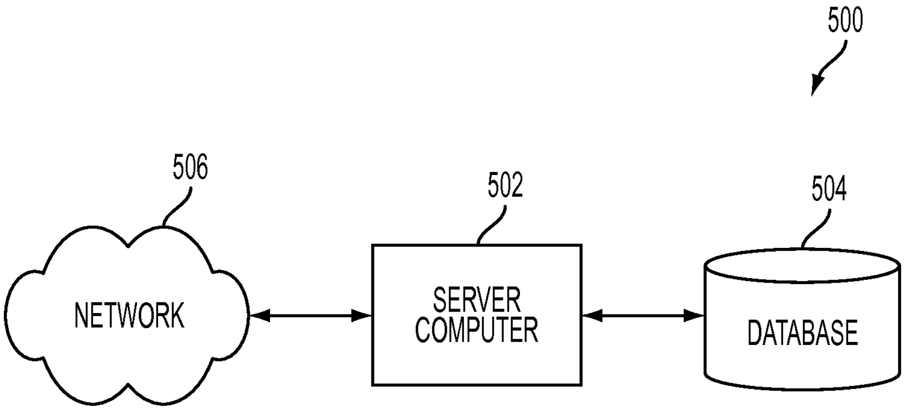

FIG. 5

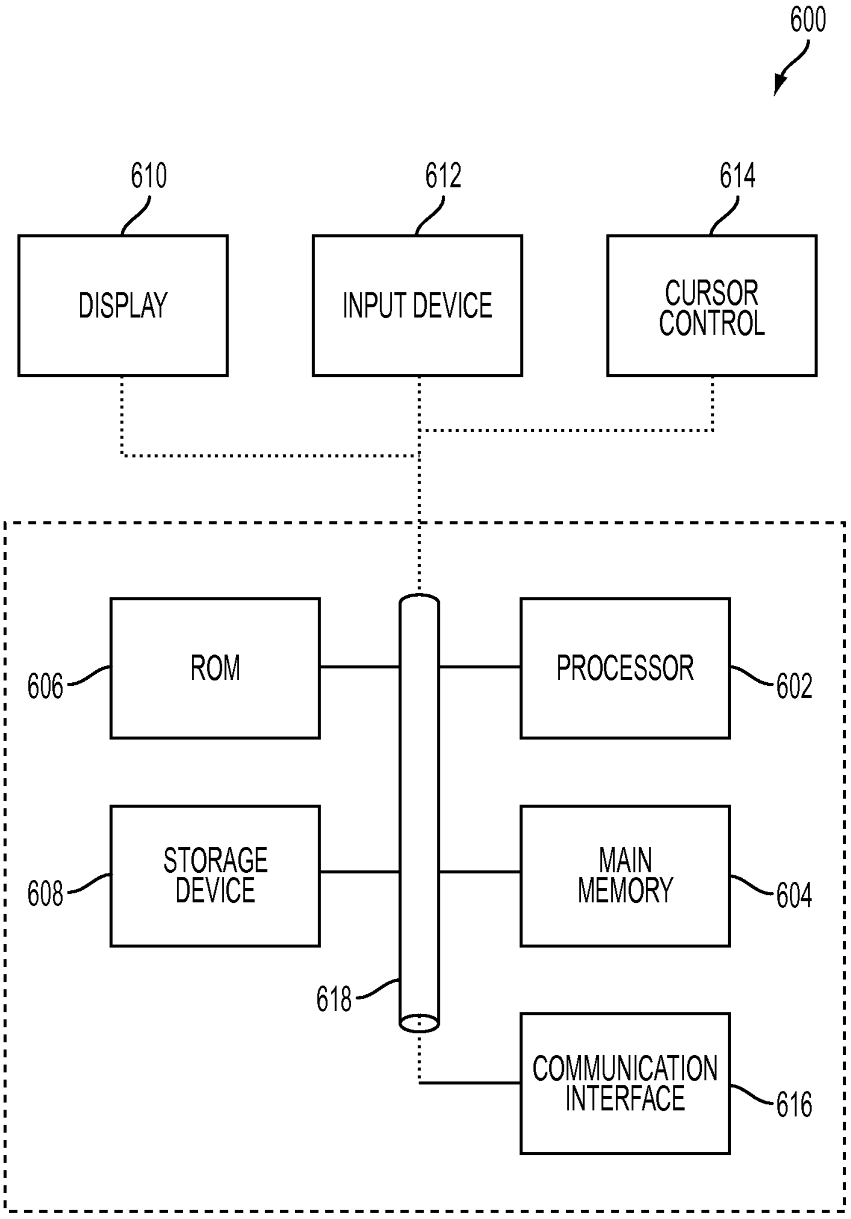
600
610
612
614
DISPLAY
INPUT DEVICE
CURSOR CONTROL
ROM
PROCESSOR
606
602
STORAGE DEVICE
MAIN MEMORY
608
604
618
COMMUNICATION INTERFACE
616

FIG. 6