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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 September 2017 (28.09.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/162997 A1

(51) International Patent Classification:
H04L 29/06 (2006.01) *H04L 12/58* (2006.01)

(21) International Application Number:
PCT/GB2017/000038

(22) International Filing Date:
23 March 2017 (23.03.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1605004.9 24 March 2016 (24.03.2016) GB

(71) Applicant: **THE SECRETARY OF STATE FOR DEFENCE** [GB/GB]; DSTL, Porton Down, Salisbury, Wiltshire SP4 0JQ (GB).

(72) Inventors: **BARNETT, Alexander, John**; DSTL IP GROUP, G02 Building 005, DSTL Porton Down, Salisbury, Wiltshire SP4 0JQ (GB). **PRESLEY, Samuel**; DSTL IP GROUP, G02 Building 005, DSTL Porton Down, Salisbury, Wiltshire SP4 0JQ (GB).

(74) Agent: **FARNSWORTH, Alastair, Graham**; DIPR Formalities Section, Poplar 2#2214, MOD Abbey Wood (South), Bristol BS34 8JH (GB).

(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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, 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, KM, ML, MR, NE, SN, TD, TG).

Published:

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

[Continued on next page]

(54) Title: A METHOD OF PROTECTING A USER FROM MESSAGES WITH LINKS TO MALICIOUS WEBSITES CONTAINING HOMOGRAPH ATTACKS

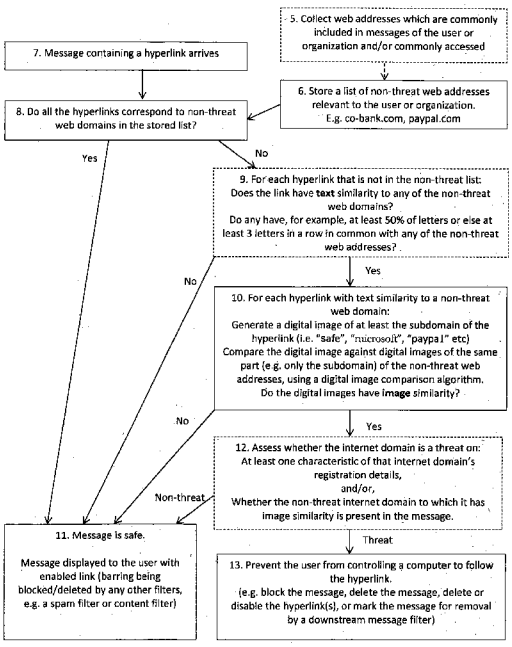

Fig 2.

(57) Abstract: Method and apparatus for protecting against homoglyph style malicious attacks by identifying threat links in incoming electronic messages such as emails. Such messages can then be blocked, deleted or edited to prevent the user controlling a computer to download code from those linked. The method involves establishing a list of internet domains are relevant to a user or organisation, and identifying incoming messages with internet links that may cause users to wrongly think those links are directed to those listed internet domains. This is achieved by applying an image similarity assessment algorithm to detect links which would have visual similarity to those listed internet domains. The method may include only assessing image similarity on those links that have a predetermined amount of text similarity to those listed internet domains, and may also include further checks if the link is still a suspect link after the image similarity check.

WO 2017/162997 A1

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

A METHOD OF PROTECTING A USER FROM MESSAGES WITH LINKS TO MALICIOUS WEBSITES CONTAINING HOMOGRAPH ATTACKS

The present invention relates to protection of computers and users from electronic messages (such as emails etc) containing links to websites which contain malware or which fraudulently purport to be a genuine service in order to collect users' personal or financial information.

Spam filters are a well-known approach for protecting users from messages containing links to websites containing malware. Techniques for filtering the messages include comparing them to templates, statistical techniques, and checking for links to websites that are known to contain malware. Some spam filters go further in looking for messages that are unusual as compared to the type of messages which are commonly received.

The inventors have identified some problems with known filters, which is that although they can stop a large proportion of untargeted spam, they tend to be ineffective against highly targeted malicious messages (known as "phishing, or "spear phishing" depending on how targeted the attack is) which can cause considerable harm to the user, the organisation, or their computer systems. A particular type of targeted attack of concern to the inventors is the homoglyph attack.

An example of a targeted homoglyph phishing attack would be to send a message to a member of staff of an organisation (e.g. a bank) containing a link that appears to be relevant to the staff of that organisation but with a small change that may not be immediately obvious to the recipient.

Well known examples are to replace the letter "o" with the number "0", perhaps using a font (in this case "Bell MT") where the number '0' looks more like a conventional "o", or using two letters that in a particular font seem similar to a single letter (e.g. "rn" looking like "m"). Other examples of the technique include www.paypa1.com and G00GLE.COM.

In addition to the more simple ASCII homoglyph attacks discussed, above it is also possible to design homoglyph attacks based on mixtures of characters from different languages. Fortunately most domain name registrars now restrict the ability to register domains with a letter from a foreign alphabet, however not all do and also it is possible to construct an attack using characters of only one language, that looks like a well-known web-domain in another language. As an example it would be possible to use the crylic letters "e" "p" and "o" (which look identical to the latin letters "e" "p" and "o") to register a domain which looks identical to the well-known website www.epo.org

The inventors have identified a weakness in known message filters, which is that in order to become widely used they need to provide generic protection for all types of users, yet the difficulty with phishing attacks is that they are tailored to be of interest to users within an organisation (or a specific individual), and thus tend to be designed to be very similar to the type of messages that the user(s) would not be surprised to received. Thus, automated methods of finding unusual messages do not tend to provide adequate protection against targeted or highly targeted phishing attacks.

Another weakness is that as compared to generic spam, with a targeted phishing attack the malware hosting website is often only registered shortly before sending of the malicious email to the victims being targeted, as this ensures that the malware hosting website will not be a known threat website at the time of performing the attack.

It would also be possible to generate a list of web domains that are visually similar to web addresses that would be familiar to users in a particular organisation. However the difficulty is that there are likely to be a vast number of such domains, taking into account that web addresses can be displayed in numerous fonts or using special language characters. As a manual task this would be so challenging as to limit both its adoption and efficacy, especially given that web addresses are displayed differently depending on the font and formatting of the email (e.g. in html) and that web addresses can use characters from many different languages which may appear similar in some fonts to characters in the language of the rest of the email text.

It is an object of the present invention to provide an apparatus and method for preventing a user from controlling a computer to access a link in a message which is likely to cause harm to the user or their organisation, or their computer systems.

According to a first aspect of the present invention there is provided apparatus for protecting a user against phishing attacks by identifying threat links in incoming electronic messages, and preventing a user from controlling a computer to follow such links, the apparatus comprising:

- An internet-domain manager operable to store a list of non-threat internet-domains that are relevant to the user or an organisation of which the user is a member, each internet-domain comprising at least a second level domain of an internet address;
- A content security manager operable to:
    - Identify suspect internet-domains in hyperlinks in the incoming electronic messages,

- Assess whether at least one of the suspect internet-domains is a threat;
- A content filter operable to directly or indirectly prevent the user from controlling a computer via a message client to follow such hyperlinks if they contain an internet-domain that has been assessed to be a threat;

characterised in that the aforesaid assessment comprises:

- providing a digital image of the suspect internet-domain;
- providing a digital image of at least one listed non-threat web address;
- comparing the two digital images by applying a digital image similarity assessment algorithm to output a measure of digital image similarity; and
- basing the assessment at least in part on that output.

This has the advantage of offering protection even if a threat internet-domain was registered only very shortly before being used in a phishing attack. It also has the advantage of addressing all types of variations in formatting and style of the message content, and the language of the characters used in the internet-domain in the message.

Preferably the assessment further comprises assessing at least one characteristic of internet-domain registration information associated with the internet-domain, against at least one criterion, and basing the assessment at least in part on that output, at least in the case that the output of the digital image similarity assessment algorithm is above a predetermined threshold.

This has the advantage of offering improved reliability, such that internet-domains with image similarity to relevant non-threat internet-domain may be passed as being non-threat if appropriate. It also enables the threshold of the digital image similarity assessment algorithm to be lower than otherwise possible (i.e. to detect more potential threats), without causing the apparatus to falsely detect too many positives.

A typical characteristic to be used is the date of registration of the internet-domain, or more specifically how recent the registration was. A recent registration may be defined as an indicator that the internet-domain is a threat web address. A historical registration may be an indicator of the internet-domain not being a threat, however other factors may be taken into account. In the absence of the use of other factors this would indicate that the internet-domain is not a threat. However the assessment may comprise other checks (either subsequent or in combination), to determine whether the internet-domain is a threat or non-threat.

Preferably the internet-domain manager is operable to collect a list of internet-domains that are commonly included in a set of electronic messages of the aforesaid organisation, and not collect in that list those internet-domains that are rarely included in the set, where at least one criterion is provided to distinguish common from rare, the collected list of internet-domains being stored as the list of non-threat internet-domains.

A simple example might be that an internet-domain is rarely included if it is included in less than one in a thousand of the messages, but a wide variety of very complex criterion could also be readily implemented by the skilled person.

For example the criteria could favour collection of internet-domains that occur rarely in electronic messages of the organisation as a whole, but are disproportionately present in messages sent to (and/or from) a minority of the users (i.e. message addresses) of the organisation.

Preferably the aforesaid collection of a list of internet-domains from the electronic messages of the organisation is performed using messages sent from message addresses controlled by the organisation (i.e. sent by members of the organisation such as the organisation's staff).

Alternatively or additionally, the internet-domain manager is operable to automatically collect a list of internet-domains that are commonly accessed by users of the organisation via web browsers of computers controlled by the organisation, the collected list of commonly accessed web addresses being stored in the list of non-threat web addresses.

Automated collection of commonly included and/or commonly accessed web addresses has the advantage that the user, or a superuser, need not manually collate this list of web addresses which provides a lower barrier to adoption by an organisation. It also typically provides for more reliable inclusion of all commonly referenced web addresses as compared to manual collection.

The content filter is operable to directly or indirectly prevent the user from controlling a computer running the message client to follow such hyperlinks. This can be achieved in several ways:

- Marking the message as malicious, and passing it to a downstream filter operable to block it, so as to not be delivered to the target user's message client.
- Deleting the message.

- Blocking the message, by not forwarding it to the target user's inbox.
- Deleting the threat link from the message.

Optionally the content filter blocks or deletes messages containing threat hyperlinks. This is the simplest solution, however alternatively the message may be forwarded with a deactivated or deleted link, typically with a warning to the user regarding why the link has been removed or deactivated.

Typically the messages are either emails or short message service messages (known as "text messages" or "sms's") or both, however the method is equally applicable to other electronic message services, including bespoke messenger services, especially those suited to being sent from one organisation to another, and especially those which typically would include links, or which are typically used in conjunction with a message client that supports the use of hyperlinks to control an internet browser.

Optionally the internet-domain manager is operable to store a list of non-threat internet-domains that are relevant to an organisation of which the user is a member. Application of the invention to an organisation rather than just an individual provides strong protection against some types of phishing attacks because attackers often target at the level of an organisation. Alternatively the list of non-threat internet-domains may be ones identified specifically with respect to the user. This provides enhanced protection against the most highly targeted phishing attacks. Optionally both organisation-level and user-level lists of relevant internet-domains are collected and used.

Advantageously, the step of assessing whether at least one of the suspect internet-domains is a threat, in the event that an internet domain of a hyperlink in a message is assessed to have image similarity to one of the non-threat internet domains, the presence of at least a sub-domain of that non-threat internet domain within the message is determined, and if that non-threat interned domain subdomain is identified within the message this is used to contribute towards a finding that the message is a threat.

This feature is advantageous because typically a phishing attack uses the name of the genuine website, often repeatedly, then including a link to similar-looking domain. For example an email falsely purporting to be from the company Paypal Inc may repeatedly mention the term "Paypal" before including a link to, for example Paypa1.com . Therefore having identified that the term

"paypa1" has image similarity to one of the trusted domains or subdomains ("paypal") the level of confidence that the link is malicious is increased by the determination that the message contains the text "paypal" (irrespective of capitalisation). This may be implemented in different ways, for example the presence of the text in the message may be used as a final check. Alternatively the required threshold for image similarity may be varied according to whether the trusted domain text is also in the message (if the text is present then a low threshold is used. If the text is absent then a high threshold is used). Alternatively checking for the text in the message may be an additional check in combination with other checks, such as checking whether the linked-to internet domain was registered only recently.

More generally therefore, in the step of assessing whether an internet-domain of a hyperlink in a message is a threat, account is taken of whether any non-threat internet domains to which the hyperlinked internet domain has image similarity to, is also present in the text of the message. Optionally account is taken of how many times it is present, whether it is present as a word rather than a text string inside another word, and/or the number of times it is present relative to the number of words in the message.

According to a second aspect of the present invention there is provided a method for protecting a user against phishing attacks by identifying threat links in incoming electronic messages, and preventing a user from controlling a computer to follow such links, the method comprising the steps of:

- Storing a list of non-threat internet-domains that are relevant to the user or an organisation of which the user is a member, each internet-domain comprising at least a second level domain of an internet address;
- Identifying suspect internet-domains in hyperlinks in the incoming electronic messages;
- Assessing whether at least one of the suspect internet-domains is a threat;
- A content filter operable to directly or indirectly prevent the user from controlling a computer via a message client to follow such hyperlinks if they contain an internet-domain that has been assessed to be a threat;

characterised in that the aforesaid assessment comprises:

- providing a digital image of the suspect internet-domain;
- providing a digital image of at least one listed non-threat web address;
- comparing the two digital images by applying a digital image similarity assessment algorithm to output a measure of digital image similarity; and

- basing the assessment at least in part on that output.

According to a third aspect of the present invention there is provided a computer program operable to control a computer to perform the method of the second aspect. Such computer program typically is recorded on a physical computer readable medium.

A preferred embodiment of the invention will now be described in detail with reference to the figures in which:

Figure 1 is an illustration of a method of preventing a user from controlling a computer to access a malicious link according to the prior art;

Figure 2 is an illustration of a method of preventing a user from controlling a computer to access a malicious link according to one embodiment of the present invention;

Figure 3 is a block diagram of a computer apparatus for protecting against phishing attacks according to a first embodiment; and

Figure 4 is a block diagram of a computer apparatus for protecting against phishing attacks according to a second embodiment.

Turning to figure 1, a message filter such as that of the prior art operates by receiving messages (step 1), identifying whether any hyperlinks in a received message are on a blacklist (step 2) and if so, blocking the message, or if not then allowing the message and active hyperlink to be passed to and displayed to the user.

Figure 2 illustrates a preferred embodiment of the present invention. A list of non-threat internet domains is stored (step 6). Each internet-domain must include at least a second-level domain (e.g. "paypal" or "hmrc"). In a typical example each internet-domain will also include the top level domain (e.g. ".com" or ".gov.uk"). Any relevant third level domain (and possibly even a file extension) could be stored too but typically these are not stored, especially not the file extension.

The list of non-threat internet domains can optionally be collected automatically by evaluating a set of messages and/or monitoring a flow of messages relevant to the organisation (step 5). Monitoring inbound messages is typically acceptable because targeted phishing attacks will generally only be a very small proportion of an inbound flow of messages. Monitoring outbound messages may however be preferred as of the small number of targeted phishing attack messages received only a small proportion of them will be replied to or forwarded by recipients within the organisation.

However the disadvantage of only monitoring outbound messages is that this may cause some internet domains that are commonly received to be omitted from the non-threat list due to the messages being of a nature which does not require a reply (e.g. emails from "noreply@..." addresses).

Optionally therefore, both inbound and outbound messages should be monitored, and preferably a higher threshold is set for inbound messages as compared to outbound messages (for example internet domains are deemed to be non-threat, if they are referenced in over 1 in 100 inbound messages, or in over 1 in 1000 outbound messages).

Optionally, at least one additional criterion is used, for example the consistency of the rate at which an internet domain is referenced in messages. Each internet domain will occur in messages a varying number of times each day (or each weekday, week, month or other time period) and this variation may follow the 'normal distribution' or may follow a skewed distribution (a criterion would be needed to distinguish normal and skewed distributions, of which many could be chosen). For internet domains which occur in a normal distribution the threshold could be lower (e.g. 1 in 2000) than for internet domains which occur in a skewed distribution (e.g. 1 in 500).

Additionally, it may be beneficial to compare the rate at which each internet domain is referenced in inbound and outbound messages, to identify the ratio between the rate of outbound references to inbound references. Phishing attacks are less likely to be responded to compared to non-threat messages, and so a lower rate would be expected. Preferably a criterion is defined to distinguish internet domains that tend to be in messages that recipients tend to reply to (and/or forward), and internet domains that tend to be in messages that recipients by comparison do not tend to reply to (and/or forward). For simplicity, this can be based only on replies (or forwards) which include the original message text, however the alternative is also possible.

It should be noted that automatic harvesting of non-threat internet domains can be performed initially based on a dataset of historical messages relevant to the organisation, or the automatic harvesting can be performed for an initial period of time on messages as they flow through (in and/or out of) the organisation.

Although the list of non-threat internet domains can be left static, it is equally possible to continuously monitor electronic messages to keep the non-threat list of internet domains up to date,

however especially in the latter case it is important to ensure that a sudden flood of phishing attacks will not cause the phishing attack internet domain to be added to the list of non-threat internet domains. This is best achieved by only (or preferentially) adding internet domains to the non-threat list if it is identified that recipients within the organisation have a high tendency to reply to messages with hyperlinks to such internet domains.

Ideally, there are provided recipient-specific lists of non-threat internet domains. This can be best achieved by applying the above methods to identify internet domains which are commonly contained in messages sent to a particular recipient within the organisation. This accounts for the fact that different users are likely to have subscribed to different newsletters or other message services. The use of user specific lists helps to provide stronger protection against phishing messages targeted at specific users. One way to generate a user specific list is to automatically identifying web domains that commonly occur in messages sent particularly to that user, either on a historical dataset (e.g. their inbox) or on messages as they pass to the user and/or updated on a continuous basis. Of course care should be taken to protect such lists as they may contain private information.

While the distinction between a commonly included internet domain and a rarely included internet domain has been described as relying on one or more criterion, in practice it would be typical for the skilled person to implement a mathematical formula for distinguishing threat and non-threat web domains in email messages of the organisation, by reference to any of the numerous statistical techniques available in the scientific literature.

Having established which internet domains are commonly included in messages and are not threats, messages are received (or continue to be received) at the organisation (step 7). The messages typically are emails (but may additionally/alternatively be text messages also known as sms's, and additionally/alternatively may be another type of electronic message).

As a general rule if the message contains no links to internet domains, or if all the links are to non-threat internet domains, then the message is not blocked/deleted/disabled. The most efficient option is to first determine that all links in a message are non-threat ones, and to bypass further checks if this is the case. Optionally the method is limited to active links in html format. However usefully the method also includes identifying any plain text that would link to a website if it were copied and pasted into a browser (i.e. web addresses), treating this text as a hyperlink, and assessing that hyperlink as described above. Such plain text can be identified by the inclusion of a "."

separating two sections of characters, the latter of which is a known top level internet domain (e.g. "com").

The described system is focussed on protecting against phishing attacks using homodyne links. Clearly therefore, it can be implemented in conjunction with other filters, such as a spam filter and/or a content filter, so even if the described system does not block/delete/disable the message, the message might still get blocked/deleted/disabled at some point for other reasons.

That said, when identifying a link in a message as corresponding to a listed non-threat internet domain, it is important to also check the top level domain.

By contrast if the message does contain links to internet domains that are not in the non-threat list (step 8), then these need to be checked.

An optional first check (step 9) is to check whether there is text similarity between the internet domain of the link in the message, and any of the internet domains in the non-threat list. Clearly oohay.com is not at risk of being confused with yahoo.com, however YAH0O.COM could be confused with YAHOO.COM. Therefore to identify web domains which could be mistaken for non-threat web domains a variety of text filters and checks could sensibly be implemented. A simple example would be to require at least 50% of the letters of the second level domain to be both in common and in the right order (in this case Y, A and H, making up 66.7% of the letters and they are found in the same order in the non-threat internet domain).

Ideally, "in common" means either being the same letter (and therefore being in the same language) or alternatively being one of a number of known homoglyphs in a list of known homoglyphs. The list for example might include:

O and 0,

l and 1,

í and l,

latin e and cryllic e,

latin p and cryllic p, and

latin o and cryllic o.

Preferably the term "homoglyph" includes typographic ligatures (situations where two letters can appear similar to a single letter). So preferably the list of known homoglyphs includes typographic ligatures. These might for example include:

rn and m,

fi and A,

cj and g,

ci and a, and

cl and d.

This makes it more difficult for an attacker to design a spoof internet domain which looks like a non-threat internet domain but lacks text similarity to it (even more difficult still if known typographic ligatures are accounted for).

Determination of text similarity can be performed in many ways. For example there are approximately 188 algorithms available on Github related to the subject of text similarity. A basic approach would be to count the proportion of letters in common (as a fraction of, for example, the average number of letters in the two text strings). One of the examples listed on Github appears to use convolutional neural networks. Another attractive option is to treat the letters as vectors and measure the cosine angle between the two vectors. This conveniently produces a value between 0 and 1. Suitable cosine-similarity text evaluation algorithms are available on the internet, or can readily be written by the PSA.

It may also be desirable to vary the threshold according to the length of one of the text strings (e.g. second level internet domains), so with long text strings (10+ characters) a high degree of text similarity would be required (high threshold) but with short text strings (E.g. 3-5 characters) a lower degree of text similarity would be required. This can be expressed as a function so the required threshold may vary smoothly but perhaps non-linearly with number of characters. Suitable thresholds or functions can be established by the PSA readily through trial and error.

With all of these approaches it is desirable to treat known homoglyphs as equivalent for purposes of text similarity, but it is very important that they are not treated as equivalent for purposes of checking whether the text strings (internet domains or sub domains thereof) are identical. So, as an example microsoft and rnicrosoft could be treated as 100% similar, but not as identical.

Note that it is possible to treat such extreme cases (microsoft and rnicrosoft) as so clearly a suspect threat (100% similar, but not identical), that checking for visual similarity may not be necessary. Thus while checking for image similarity is a key feature of the invention, it is not essential to perform it on every suspect internet domain. Preferably however the additional step of checking the internet domain's registration information is performed, as this helps to distinguish between accidentally similar domains and deliberately similar domains.

To ensure different characters in different languages are not wrongly treated as identical they should be converted to (or more generally, provided in) a format, e.g. numbers, which distinguishes all such characters, prior to the step of assessing text similarity. Unicode (preferably the most up to date version) is a sensible option.

Preferably therefore the step of identifying suspect internet-domains in hyperlinks in the incoming electronic messages, includes first assessing a measure of text similarity between any such internet domains in hyperlinks and each of the non-threat internet domains, with respect to a criterion, and treating any internet domains in hyperlinks that meet the criterion as suspect, and any internet domains in hyperlinks that do not meet the criterion as not suspect, and the step of assessing whether at least one of the suspect internet-domains is a threat is applied to those internet domains in hyperlinks that are treated as suspect.

Preferably also the assessment of a measure of text similarity includes providing a list of known pairs of single or double character homoglyphs, and in each instance of comparing an internet domain in a hyperlink with a non-threat internet domain any of the listed homoglyphs are identified and treated as identical for the purposes of measuring text similarity.

One way to perform the text assessment while treating listed homoglyphs as identical is to convert one of the homoglyphs (particularly the one in the internet domain in the incoming message) into its counterpart prior to text similarity assessment. Another way is to convert them both into an arbitrary code.

For each such listed homoglyph it is desirable to repeat the text similarity assessment, both with and without treating the homoglyphs as identical. This is because a phishing attack may include the characters "rn" without relying on the similarity between "rn" and "m" to look similar to a domain that is well known to staff/users in that particular organisation. For example if "armorm.com" was a

website commonly accessed by an organisation, and a message arrived with a link to "arnorn.com". The font used in the email might not cause "rn" to look very similar to "m", and by replacing each "rn" with an "m" (creating amom.com) you potentially cause the text similarity assessment alrogithm to produce a lower similarity assessment output than if you had not done so. Additionally if multiple listed homoglyphs are detected, the text similarity for each permutation (changed or not changed) should be tested. Where multiple versions are tested, the highest output (strongest indication of similarity) should be selected for determining whether there is text similarity.

It is desirable that where any links in messages include a third level domain (and/or other sub-domain) then that sub-domain should be checked against the second level domains of the non-threat lists. In this case a direct match does not lead to the link being considered non-threat, but (barring other factors) will normally lead to the link being considered a phishing threat.

So for example if a message is received with a link to "bbc.iplayer.top-bbc-content-2015.co.uk" then it is desirable that "bbc" and "iplayer" are compared to the non-threat internet domains (specifically the second level domains), and a match would (barring other factors that may be taken into account, if any) lead to the link being deemed to be a phishing threat. This approach would guard against attacks where the attacker is hoping that the recipient will believe that the left hand side of the URL is the website address (E.g. "bbc.co.uk"), when in fact the website directed to is in this case a part of "top-bbc-content-2015.co.uk".

Similarly the combination of two or more subdomains in a suspect hyperlink should be considered together, as well as separately.

In step 10, any suspect hyperlinks are evaluated via digital image comparison. The second level domain of the hyperlink is converted into a digital image (i.e. generally a rectangular image, generally with data defining black areas and white areas, and typically not defining colour).
Similarly the second level domain of each of the listed non-threat internet domains is either converted to a digital image or already provided as a digital image. Generally the same method of converting to digital image should be used on both second level domains. Ideally for text formatted messages, ideally the suspect link second level domain is converted from plain text to a digital image, and also a digital image is generated of its formatted appearance, and both digital images are compared to the digital images of the non-threat second level domains. Suitable methods of

converting text to an image are easily selected and obtained (or else generated if necessary) by the skilled person.

Comparison of two digital images is performed on a computer by an image similarity assessment algorithm. Many relevant algorithms already exist and can be selected and modified to suit this task. Searching for image comparison on github for example reveals a number of algorithms and even an algorithm for comparing image comparison algorithms. Although one option is to use a simple image comparison algorithm, which simply compares whether each part of the two images are the same, it is preferable to use an algorithm which assesses the degree to which the structural features of one image are present in the other in similar but not necessarily exactly matching locations. This provides a better estimate of whether a user is at risk of mis-reading the link as being a link to a common internet domain. Such an algorithm can be written by the skilled person as an alternative to using and optionally modifying an available image comparison algorithm.

By way of example, starting from a simple image comparison algorithm, the skilled person could generate a more sophisticated algorithm to compare images A and B as follows: Firstly divide digital images A into an array of smaller digital images (optionally with extra overlapping ones) so that each small digital image will be a part of a single letter depicted in the original, and each part of each letter should be contained in at least one small digital image. Now compare each of the small digital images against a corresponding (but larger) area of image B. This may involve repeatedly trying to match the small digital image against slightly different locations of image B to find the best match (e.g. via cross correlation or phase correlation). The similarity between images A and B is output as a measure of how many of the small images from A can be readily matched to a similar location of image B.

Further, where a suspect internet domain has already been identified as having text similarity to a non-threat internet domain, then it may be preferably to only compare the parts which lack text similarity, preferably along with the any immediately neighbouring letters. So for example comparing A: "microsoft" to B: "rnicrosoft" would involve first identifying that 8 of the 9 letters in A are present in B in the correct order, and thus the subdomains have text similarity. However then it is only necessary to compare a digital image of the letters "mi" against a digital image of the letters "rni" for example as they would be displayed based on any formatting defined in the message (e.g. a specific font).

If the output of the digital image similarity algorithm indicates high image similarity then the suspect hyperlink is either deemed a threat, or remains a suspect link for further analysis (step 12). If the digital image similarity algorithm indicates low image similarity then the suspect hyperlink is deemed a non-threat and the message is passed to the user's inbox or message client for viewing without being deleted, blocked or disabled (step 11).

In a preferred embodiment a hyperlink that remains a suspect hyperlink following image similarity assessment (step 10) is evaluated further, with a further assessment being based on at least one characteristic of the relevant internet domain's registration details. The characteristic assessed is how recently the relevant domain was registered (i.e. recently or historically), based on whether this was longer ago than a threshold value. So, for example if the domain was registered less than a week prior to receipt of the message, then this in combination with the identified image similarity is used to deem the hyperlink a threat. The user is then prevented from controlling a computer to access the hyperlink address – either by deleting the message, blocking the message, deleting or disabling the hyperlink or at the least marking the message as malicious in order to cause a further filter (such as a spam filter, or content filter or other type of filter) downstream to delete or block the message.

Generally the internet domain's registration details are obtained when needed by reference to a suitable resource at a predetermined location on the internet which may be a 3rd party website or may be a proprietary server providing a bespoke service for detecting phishing attacks.

In another example an assessment is made of whether the non-threat internet domain (which has been assessed as having image similarity to an internet domain of a hyperlink) is mentioned in the text of the message. If the text string merely occurs incidentally (for example the internet domain "ample" might occur in a message which contains the word "example") this could sensibly be disregarded. However the presence of the non-threat internet domain as a word within the message would be a strong indicator that the message is malicious.

While this step has been discussed as a separate subsequent step, alternatively the steps are combined and the image similarity and one or more other criteria are assessed jointly via an algorithm which balances two or more input variables to provide an assessment of whether the message is a threat. Such algorithms may vary widely, perhaps relying on many inputs such as the number of times the word is repeated, the amount of image similarity, the recentness of the URL

registration, whether the non-threat domain is in an organisation relevant non-threat list or a personal non-threat list etc, and/or whether the non-threat internet domain is one known to provide login-based access or merely freely accessible information.

In most cases the link has been described as a hyperlink – generally this includes non-enabled or inactive hyperlinks in the form of a URL in plain text (whether or not the message itself is in plain text format). This is preferable because some phishing messages may merely invite the user to copy and paste the URL into a browser address bar.

The method and apparatus provide for protecting against homoglyph phishing attacks by identifying threat links in incoming electronic messages such as emails. Such messages can then be blocked, deleted or edited to prevent the user being duped by them. The method involves establishing a list of internet domains are relevant to a particular organisation, and identifying incoming messages with internet links that may cause users to wrongly think those links are directed to those listed internet domains. This is achieved by applying an image similarity assessment algorithm to detect links which would have visual similarity to those listed internet domains. The method may include only assessing image similarity on those links that have a predetermined amount of text similarity to those listed internet domains, and may also include further checks if the link is still a suspect link after the image similarity check.

Any suitable computing system or group of computing systems can be used for performing the operations described herein. For example, Figure 3 is a block diagram depicting an example of a computing system 14 that executes a link assessment engine 15 for assessing electronic messages as to whether they contain generating data structures that encode logical propositions as described above.

The computing system 14 includes a processor 16 that is communicatively coupled to a memory 15 and that executes computer-executable program code and/or accesses information stored in the memory 15. Examples of the processor 16 include (but are not limited to) a microprocessor, an application-specific integrated circuit ("ASIC"), a field-programmable gate array ("FPGA"), or other processing device. The processor 16 can include any number of processing devices, including one.
The memory 15 includes any suitable non-transitory computer-readable medium. In various embodiments, the computer-readable medium includes any electronic, optical, magnetic, or other

storage device capable of providing a processor with computer-readable instructions or other program code.

Non-limiting examples of a computer-readable medium include a CD-ROM, a DVD, a magnetic disk, a memory chip, a ROM, a RAM, an ASIC, optical storage, magnetic tape or other magnetic storage, or any other medium from which a computer processor can read instructions.

The computing system 14 may also include a number of external or internal devices such as input or output devices. For example, the computing system 14 is shown with an input/output ("I/O") interface 18 that can receive input from input devices or provide output to output devices. A bus 17 can also be included in the computing system 14. The bus 17 can communicatively couple one or more components of the computing system 14.

The computing system 14 executes program code that configures the processor 16 to perform one or more of the operations described above. The memory 15 stores this program code. The program code includes, for example, the link assessment engine 15 or any other suitable engine, module, or application that can be used to perform one or more operations described herein. The program code may be resident in the memory 15 or any suitable computer-readable medium and may be executed by the processor 16 or any other suitable processor. In some embodiments, the program code includes processor-specific instructions generated by a compiler and/or an interpreter from code written in any suitable computer-programming language, including, for example, C, C++, C#, Visual Basic, Java, Python, Perl, JavaScript, and ActionScript.

The computing system 14 receives electronic messages via I/O device 18 from message server 19, and typically sends them back to the same or another message server for onward transmission, or directly on to be picked up by the relevant message client. Computer 20 is typically one of many computers, typically each owned by the organisation and typically each controlled by a user who is a member of the organisation. The computer, which may be a smartphone or other portable device, 20 similarly has memory, processor, bus and an I/O device for receiving messages 21, and an I/O device which is generally a screen but may be another visual (or even audio) display unit 23 for displaying the electronic message to a user 24 (dotted arrow), and allowing the user 24 to control the computer 20 to follow a hyperlink within the displayed message (second dotted arrow).

The dotted formatting of the arrows indicate that one or both of these steps are selectively enabled or prevented. Either the user is prevented from viewing the message (by it being blocked or deleted

by the computing system 14) or the hyperlink is deactivated or removed such that the user is prevented from controlling the computer 20 from activating the link and downloading content from an internet domain 25 referenced by the link, by means of an internet connection 26.

Figure 4 shows an alternative embodiment where the Computing system 14 and computer system 20 are provided in one computer system 14. This computer system both assesses links to identify targeted phishing attacks, and also displays messages to the client via a message client. The numbering system of figure 3 is used in figure 4.

General Considerations

Numerous specific details are set forth herein to provide a thorough understanding of the claimed subject matter. However, those skilled in the art will understand that the claimed subject matter may be practiced without these specific details. In other instances, methods, apparatuses, or systems that would be known by one of ordinary skill have not been described in detail so as not to obscure claimed subject matter.

The system or systems discussed herein are not limited to any particular hardware architecture or configuration. A computing device can include any suitable arrangement of components that provides a result conditioned on one or more inputs. Suitable computing devices include multipurpose microprocessor-based computer systems accessing stored software that programs or configures the computing system from a general purpose computing apparatus to a specialized computing apparatus implementing one or more embodiments of the present subject matter. Any suitable programming, scripting, or other type of language or combinations of languages may be used to implement the teachings contained herein in software to be used in programming or configuring a computing device.

While the present subject matter has been described in detail with respect to specific embodiments thereof, it will be appreciated that those skilled in the art, upon attaining an understanding of the foregoing, may readily produce alterations to, variations of, and equivalents to such embodiments. Accordingly, it should be understood that the present disclosure has been presented for purposes of example rather than limitation, and does not preclude inclusion of such modifications, variations, and/or additions to the present subject matter as would be readily apparent to one of ordinary skill in the art.

Claims:

1. Apparatus for protecting a user against phishing attacks by identifying threat links in incoming electronic messages, and preventing a user from controlling a computer to follow such links, the apparatus comprising:

- An internet-domain manager operable to store a list of non-threat internet-domains that are relevant to the user or an organisation of which the user is a member, each internet-domain comprising at least a second level domain of an internet address;
- A content security manager operable to:
    - Identify suspect internet-domains in hyperlinks in the incoming electronic messages,
    - Assess whether at least one of the suspect internet-domains is a threat;
- A content filter operable to directly or indirectly prevent the user from controlling a computer via a message client to follow such hyperlinks if they contain an internet-domain that has been assessed to be a threat;

characterised in that the aforesaid assessment comprises:

- providing a digital image of the suspect internet-domain;
- providing a digital image of at least one listed non-threat internet-domain;
- comparing the two digital images by applying a digital image similarity assessment algorithm to output a measure of digital image similarity; and
- basing the assessment at least in part on that output.

2. Apparatus according to claim 1, wherein the assessment additionally comprises assessing at least one characteristic of internet-domain registration information associated with the internet-domain, against at least one criterion, and basing the assessment at least in part on that output, at least in the case that the output of the digital image similarity assessment algorithm is above a predetermined threshold.

3. Apparatus according to claim 2, wherein at least one characteristic used is how recent the internet-domain registration was.

4. Apparatus according to any preceding claim, wherein the internet-domain manager is operable to collect a list of internet-domains that are commonly included in a set of electronic messages of the aforesaid user or organisation, and not collect in that list those internet-domains that are rarely included in the set, where at least one criterion is provided to distinguish common from rare, the collected list of internet-domains being stored as the list of non-threat internet-domains.

5. Apparatus according to any preceding claim, wherein the internet-domain manager is operable to automatically collect a list of internet-domains that are commonly accessed by the user or users of the organisation, via web browsers of computers controlled by the user or organisation, the collected list of commonly accessed web addresses being stored in the list of non-threat web addresses.

6. Apparatus according to any preceding claim, wherein the content filter blocks or deletes messages containing threat hyperlinks.

7. Apparatus according to any preceding claim, wherein the step of identifying suspect internet-domains in hyperlinks in the incoming electronic messages, includes:

- first assessing a measure of text similarity between any such internet domains in hyperlinks and each of the non-threat internet domains, with respect to a criterion, and treating any internet domains in hyperlinks that meet the criterion as suspect, and any internet domains in hyperlinks that do not meet the criterion as not suspect, and

the step of assessing whether at least one of the suspect internet-domains is a threat is applied to those internet domains in hyperlinks that are treated as suspect.

8. Apparatus according to claim 7, wherein the assessment of a measure of text similarity includes providing a list of known pairs of single or double character homoglyphs, and in each instance of comparing an internet domain in a hyperlink with a non-threat internet domain any of the listed homoglyphs are identified and treated as identical for the purposes of measuring text similarity.

9. Apparatus according to any preceding claim, wherein the internet-domain manager is operable to store a list of non-threat internet-domains that are relevant to an organisation of which the user is a member.

10. Apparatus according to any preceding claim, wherein in the step of assessing whether an internet-domain of a hyperlink in a message is a threat, account is taken of whether any non-threat internet domains to which the hyperlinked internet domain has image similarity to, is also present in the text of the message.

11. A method for protecting a user against phishing attacks by identifying threat links in incoming electronic messages, and preventing a user from controlling a computer to follow such links, the method comprising the steps of:

- Storing a list of non-threat internet-domains that are relevant to the user or an organisation of which the user is a member, each internet-domain comprising at least a second level domain of an internet address;
- Identifying suspect internet-domains in hyperlinks in the incoming electronic messages;
- Assessing whether at least one of the suspect internet-domains is a threat;
- A content filter operable to directly or indirectly prevent the user from controlling a computer via a message client to follow such hyperlinks if they contain an internet-domain that has been assessed to be a threat;

characterised in that the aforesaid assessment comprises:

- providing a digital image of the suspect internet-domain;
- providing a digital image of at least one listed non-threat internet-domain;
- comparing the two digital images by applying a digital image similarity assessment algorithm to output a measure of digital image similarity; and
- basing the assessment at least in part on that output.

12. A computer program operable to control a computer to perform the method of claim 11.

13. Apparatus or method substantially as hereinbefore described with reference to figures 2 to 4.

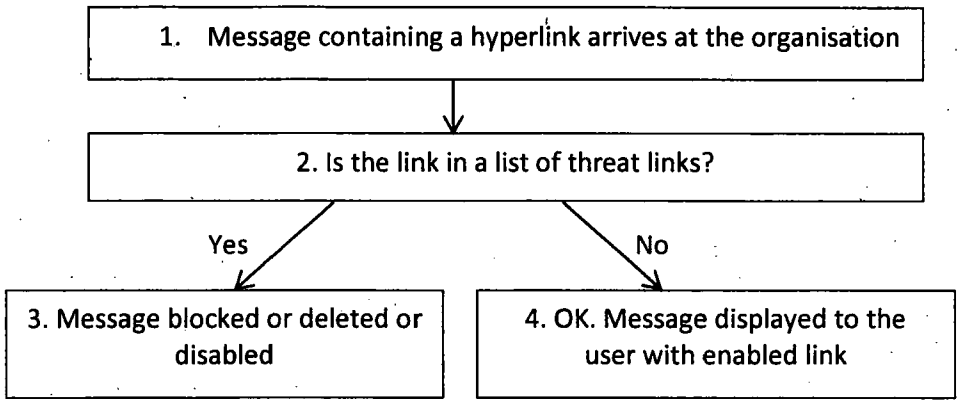

Fig 1. Prior Art

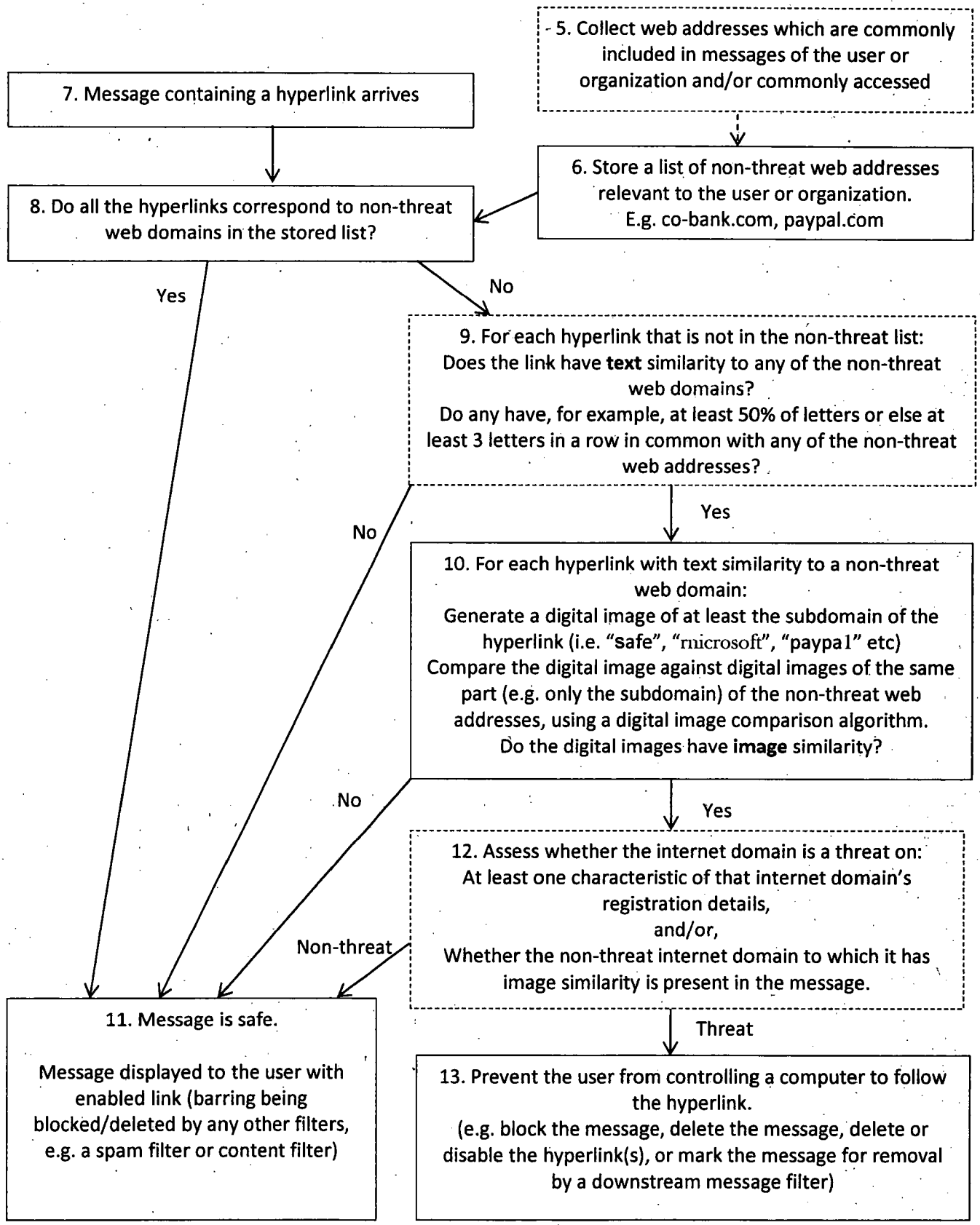
5. Collect web addresses which are commonly included in messages of the user or organization and/or commonly accessed
6. Store a list of non-threat web addresses relevant to the user or organization. E.g. co-bank.com, paypal.com
7. Message containing a hyperlink arrives
8. Do all the hyperlinks correspond to non-threat web domains in the stored list?
Yes
No
9. For each hyperlink that is not in the non-threat list: Does the link have text similarity to any of the non-threat web domains? Do any have, for example, at least 50% of letters or else at least 3 letters in a row in common with any of the non-threat web addresses?
Yes
No
10. For each hyperlink with text similarity to a non-threat web domain: Generate a digital image of at least the subdomain of the hyperlink (i.e. "safe", "rnicrosoft", "paypa1" etc) Compare the digital image against digital images of the same part (e.g. only the subdomain) of the non-threat web addresses, using a digital image comparison algorithm. Do the digital images have image similarity?
Yes
No
12. Assess whether the internet domain is a threat on: At least one characteristic of that internet domain's registration details, and/or, Whether the non-threat internet domain to which it has image similarity is present in the message.
Threat
Non-threat
11. Message is safe. Message displayed to the user with enabled link (barring being blocked/deleted by any other filters, e.g. a spam filter or content filter)
13. Prevent the user from controlling a computer to follow the hyperlink. (e.g. block the message, delete the message, delete or disable the hyperlink(s), or mark the message for removal by a downstream message filter)

Fig 2.

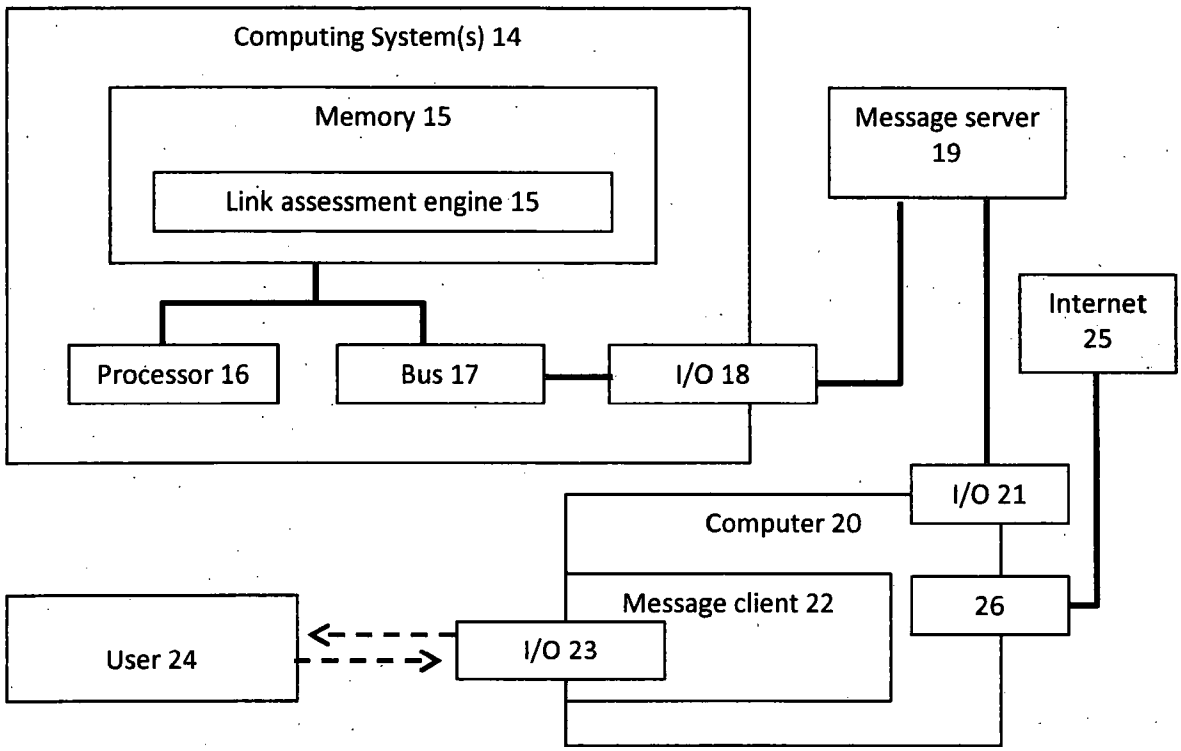
Computing System(s) 14
Memory 15
Link assessment engine 15
Processor 16
Bus 17
I/O 18
Message server 19
Internet 25
I/O 21
Computer 20
Message client 22
26
User 24
I/O 23

Fig 3.

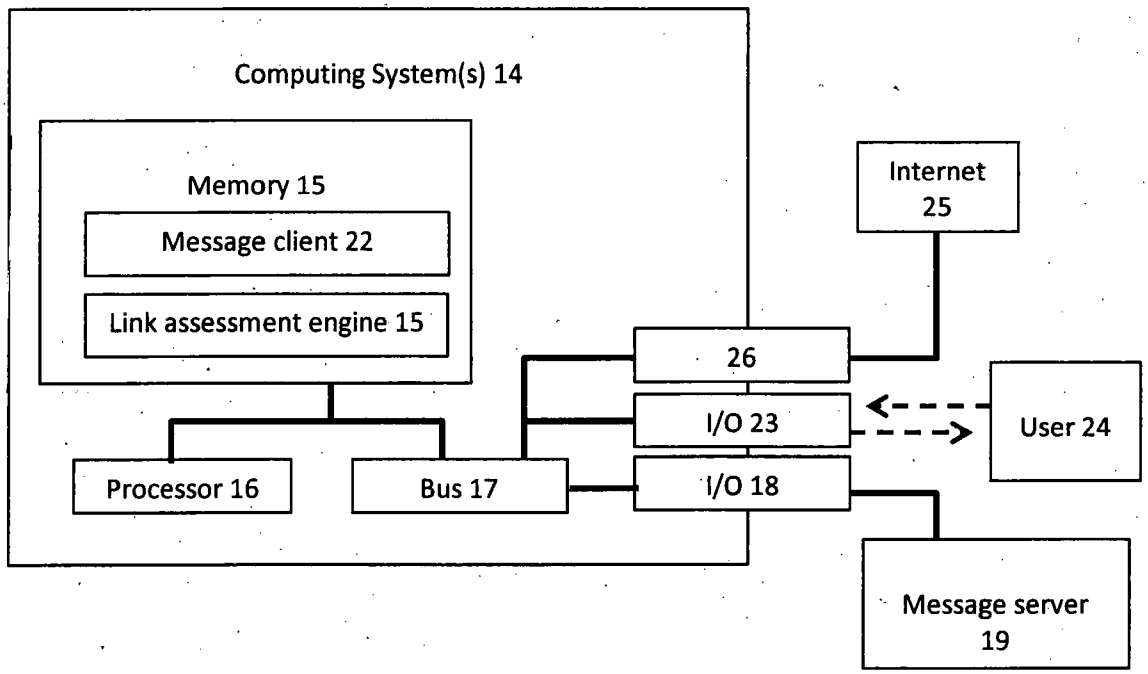
Computing System(s) 14
Memory 15
Message client 22
Link assessment engine 15
Processor 16
Bus 17
Internet 25
26
I/O 23
I/O 18
User 24
Message server 19

Fig 4.

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2017/000038

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L29/06 H04L12/58
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2008/086924 A1 (IBM [US]; IBM UK [GB]; REUMANN JOHN [US]; VERMA DINESH [US]) 24 July 2008 (2008-07-24)	1,4-13
Y	abstract; figures 1-4 page 2, line 6 - page 3, line 16 page 4, line 4 - page 5, line 7 page 7, line 4 - page 9, line 2; figure 2 page 11, line 3 - page 11, line 4	2,3
Y	CN 105 357 221 A (BEIJING QIHOO TECHNOLOGY CO; QIZHI SOFTWARE BEIJING CO LTD) 24 February 2016 (2016-02-24)	2,3
A	abstract paragraph [0027] - paragraph [0027]	1

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 May 2017	17/07/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Jakob, Gregor

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2017/000038

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2008086924 A1	24-07-2008	AT 497620 T	15-02-2011
		CA 2673322 A1	24-07-2008
		EP 2104901 A1	30-09-2009
		ES 2359466 T3	23-05-2011
		IL 200487 A	30-05-2013
		JP 2010516007 A	13-05-2010
		KR 20090108000 A	14-10-2009
		US 2008172741 A1	17-07-2008
		US 2012304295 A1	29-11-2012
		WO 2008086924 A1	24-07-2008
CN 105357221 A	24-02-2016	NONE	