The Phorm "Webwise" System

Richard Clayton, 4th April 2008

Introduction

On Wednesday 26th March 2008 I put on my "hat" as Treasurer of FIPR and accompanied Becky Hogge, Director of the Open Rights Group, to a meeting with Phorm, the company whose advertising platform has been much in the news lately. We had a wide-ranging briefing about the technical aspects of their system, the way in which it preserves privacy and their vision of how it will transform online advertising. The meeting was very technical throughout, and these notes reflect that. Doubtless more user-friendly explanations will be provided in time.

The meeting was entirely on the record, with the sole agreed exception of Phorm telling us the identities of their suppliers of "phishing" URLs. They fully understood that we would be writing about what we learnt from the meeting. Some of the information which was imparted had already been made public in various places, but a fair bit is entirely new.

I provided an initial draft of this document to Phorm, who corrected a handful of details (mainly timeout periods) which they'd mis-remembered in the meeting, and pointed out a couple of errors I'd made – which I have been happy to correct. Because the relevant person was travelling, this has engendered a slight delay, but I felt that the resulting accuracy of the document made this worthwhile.

Naturally, when describing such a complex system, there is ample opportunity for errors to creep in, or important detail to have been glossed over. A single meeting is inadequate for delving into every detail – hence it must be assumed that there is still more to be usefully learnt about the system's operation.

I have tried to stay away from commentary as to the social, moral or legal issues that arise from the system, but to concentrate on the technicalities of its operation.

Finally, and very importantly, nothing within these notes should be taken as any sort of approval or endorsement of Phorm's system design or implementation. It is merely a record of my current understanding of how their system works.

As a road-map to what's in these notes, I shall be discussing:

- the basic traffic inspection architecture;
- how Phorm's tracking cookies are handled;
- which web requests and web pages are inspected;
- what data from this is processed by Phorm;
- how they map your browsing into advertising "channels";
- the advert serving mechanism;
- and, various other matters that don't fit neatly into other sections.

A Inspecting traffic

1. The basic concept behind the Phorm architecture is that they wish to take a copy of the traffic that passes between an end-user and a website. This enables their systems to inspect what requests were made to the website and to determine what content came back from that website. An understanding of the types of websites visited is used to target adverts at particular users.

2. The actual mechanics of taking the copy differs from ISP to ISP, but one can view it as a "Layer 7 switch", implemented using Policy Based Routing (PBR) or Deep Packet Inspection (DPI). This switch is capable of providing a view of the web session to out-of-band machines. By "out-of-band" I mean the original session is not affected by the act of making a copy, and neither end is capable of directly determining that a copy has been made.

3. The "Layer 7 switch" only inspects traffic on port 80, the conventional port used for web browsing using the HTTP protocol. Traffic on other ports will be entirely ignored by the Phorm system.

4. Since the device is a Layer 7 switch, it understands the HTTP protocol itself, and can pick apart the requests and responses that are being made. If the traffic does not appear to be HTTP (it is another protocol using port 80, or perhaps it is encrypted) then the traffic will be ignored by the Phorm system.

5. The "Layer 7 switch" is also capable of redirecting traffic so that it does not reach the "true destination" but instead is serviced by a machine within the ISP's network that, for example, does some sleight-of-hand to check whether the user has opted-out of the system, and if not, to determine the Unique Identifier (UID) by which they are known to the Phorm system.

6. The various ISPs who will implement the Phorm system may operate their own opt-in or opt-out systems, these are not considered further.

7. In the meeting, Phorm stated their preference for an opt-out system, indicating that they believed this would lead to higher overall usage.

B Cookies

8. A quick review of "cookie" handling is order... more details can be found in RFC2695 and the original Netscape specification.

9. Along with requested content, a website can return a text string called a "cookie" which will be automatically stored by the user's browser. Any further requests to the same website will be automatically accompanied by the cookie, and hence the website will be able to link requests together – perhaps to record progress through a procedure, or to keep track of visitor preferences.

10. It is a key design aspect of cookies that they are linked to a particular domain and are only returned to a website within that domain – which means that a given website will never receive a cookie that is associated with a completely different website.

11. However, cookies can be supplied by a website with the name of another website within them. This is often associated with banner ads, served from a different domain. This form of cookie is called a "third party" cookie and modern browsers will disable them (viz: the cookies are not stored or returned) if the user requests this.

12. Further discussion will be solely about "first party" cookies which are only set by and returned to a particular domain. These cookies can also be disabled on user request either for a particular site or for all sites, although in practice many sites do not work well without cookies being enabled.

13. Turning now to the Phorm system. Consider the first web request made by a user, for, let us say, `http://www.cnn.com/index.html`. This will take the form of a `GET` request for `index.html` with a `HOST` header of `www.cnn.com`.

14. The Layer 7 switch will see that the request does not contain a Phorm "cookie" and will direct the request to a machine located within the ISP network that will pretend to be `www.cnn.com` and will return a "`307`" response which says, in effect, "you want that page over there". The page that will be directed to is `webwise.net/bind/?<parameters>` where the parameters record the original URL that was wanted.

15. The user's browser will now wish to visit the `webwise.net` page it has been redirected to, and will issue an appropriate `GET` request for this page. If the user already has a cookie for `webwise.net` then this will, as is standard, accompany the request.

16. The Layer 7 switch will again direct the request to a special machine (within the ISP's network for performance reasons if nothing else). This special machine, which is now acting as `webwise.net`, will inspect any existing cookie to establish the current UID associated with the user. If there is no cookie then a new UID will be issued instead.

17. The response from `webwise.net` will be a `307` response redirecting the user to a special URL on `www.cnn.com`. The response will also contain a cookie (in the `webwise.net` domain) which contains the UID that is used to track the user. The special URL will also contain a copy of this UID, along with the original request that the user made.

18. The special URL on `www.cnn.com` will now be fetched by the user's browser, and the Layer 7 switch will recognise the request (from its form) as once again to be redirected to the special machine, which will once again pretend to be `www.cnn.com`.

19. The special machine will return a third and final `307` redirection, and this time the destination URL will be the `www.cnn.com/index.html` page that the user has been waiting to visit all along.

20. The response in paragraph 19 will also set a special "`webwise`" labelled cookie within the `www.cnn.com` domain – which it will expect to be accepted because the machine is pretending to be `www.cnn.com`. This cookie will contain the user's UID.

21. Finally the user's browser will re-issue the original request for `www.cnn.com/index.html` but this time it will be accompanied by the `webwise` cookie that has just been set in the `www.cnn.com` domain, and so the Layer 7 switch will permit it to pass through to the real CNN site.

22. The specious cookie (from the point of view of `www.cnn.com`) will be removed as the request passes through the Layer 7 switch.

23. The cookie has a lifetime of three days.

24. If, later on, the `www.cnn.com` website was to be visited via another ISP that was not using a Phorm system (or if subsequent accesses were made using the "`https`" protocol) then the cookie would reach `www.cnn.com`.

25. Phorm believe that by placing their name (`webwise`) within the cookie they place within the `www.cnn.com` domain, no clash – or other bad effects – can occur.

26. Further requests for `www.cnn.com` pages (and all the other bits and pieces that make up a modern web page, such as images, pop-ups, cascading style sheets and so on) will automatically contain the `webwise` cookie, and so there will be no need to redirect any of these. The behaviour at this stage underlies the claim made by Phorm that their system does not slow down browsing.

27. If the user has disabled cookies for CNN (viz: they don't record their values and don't supply them with further requests), then there is potential for an infinite loop – repeating all the `307` responses forever. The Layer 7 switch recognises this situation and records that future traffic (at least for a while) from the particular IP address to the particular (CNN) domain is not to be redirected.

28. If the user has set a cookie within the `webwise.net` domain indicating that they do not wish to be tracked, then this preference is passed to the Layer 7 switch during the process in paragraph 16 above. The details on how this is done were not explained by Phorm... but it is presumably related to the mechanism described in the previous paragraph.

29. If the user does not accept any cookies in the `webwise.net` domain then they will always be allocated a new identifier for every website they visit. This situation is detected by the Layer 7 switch and the IP address is "blacklisted" and future traffic is not redirected.

30. Note that the blacklisting of IP addresses by the Layer 7 switch (as described in the three previous paragraphs), whether general, or for particular domains, will apply to all of the users who are sharing a particular IP address, not just users with a particular UID. However, because the "blacklisting" will time out eventually, the exact behaviour will depend upon the mixture of requests made by different users who have different browser settings.

31. Phorm told us that the UID which is allocated to the user is a 16 byte value chosen at random. That is to say it is just a number. It is not, for example, an encryption of some data that might later be decrypted. The actual value sent on the wire will be base-64 encoded, so it will be seen by humans as a 22 character string.

32. If, for whatever reason, the user discards cookies for other websites, such as `www.cnn.com`, then – provided that they have not discarded their `webwise.net` cookie – they will retain their existing UID.

33. If the user discards their `webwise.net` cookie then they will be continue to be tracked under their old identity for up to three days whilst visiting sites that they have visited before (because of the cookie in that website's domain). They will however acquire a new UID for all new websites. Phorm have no way of linking the old UID and the new UID together, so the user in effect gets a fresh new identity.

C Web pages that are inspected

34. It should now be clear exactly how Phorm's UIDs come to be included within cookies in all requests made by the user's browser. I now report upon the way in which the Phorm system inspects the browsing activity to keep track of the web pages that the user is viewing. Recall from paragraph 2 that the Layer 7 switch mirrors a copy of all of the browsing activity to an out-of-band machine, it is this machine that inspects the traffic.

35. In order to reduce the workload of later stages, only some of the web traffic gets fully examined. A number of filetype extensions (such as for images) are completely ignored, and only pages with a "`Content-Type`" of "`text/html`" are further processed.

36. To avoid processing non-web traffic, the Phorm system has a "whitelist" of "`User-Agent`" identification strings, the type and version text that browsers place into their requests. If an HTTP request does not appear to have been generated by a "well-known" browser, then the request will be ignored.

37. Sites that use "basic auth" (RFC 1945 et seq.) will be ignored, viz: sites where the browser remembers a user name and password and supplies it with each request to avoid the need to log in again.

38. If the website is on a blacklist list of "webmail" sites, viz: sites where people can read or compose private email, then the traffic will be ignored. This list currently contains "more than 25 sites".

39. When a website is first visited (by any ISP customer) the pages are not inspected. Instead, a request is queued to fetch the site's "`robots.txt`" file; viz: a file maintained by the website owner which tells web crawlers and other automated systems which parts of the website should not be indexed or processed.

40. Once the `robots.txt` file (if any) has been fetched, it will be cached. The cache retention period will be value set by the website using standard HTTP cache-control mechanisms, or for one month if no period is specified. The minimum period that the file will be cached for is two hours.

41. The `robots.txt` file will be inspected and URLs that fall within forbidden areas of the website will not be processed by the Phorm system.

42. This mechanism, which will permit website owners to opt their pages out of the Phorm system, does not seem to have been previously described in any of Phorm's documentation. They were unable to provide an explanation as to why this had not previously been disclosed.

43. In the meeting, Phorm were unable to tell us the `User-Agent` string they match against in the `robots.txt` file, knowledge of which would be required if a website owner wished to set particular rules for Phorm which differed from, for example, for the GoogleBot.

44. I asked for further clarification and was told "we work on the basis that if a site allows spidering of its contents by search engines, then its material is being openly published. Conversely, if the site has disallowed spidering and indexing by search engines, we respect those restrictions in `robots.txt`".

45. It therefore still remains unclear to me what the Phorm system does if the robots.txt file does not use a `User-Agent: *` construction, and whether this will be in line with what the website owner intended.

D Determining what type of pages are being visited

46. If the web page request is made for the search page of a major search engine then the search terms, which are encoded within the URL, will be parsed and extracted.

47. The system will associate the web page that is returned from any website with the original request. It will parse this webpage, unless it is the result of a `POST`.

48. The page is broken up into individual "words". Words which are solely made up of digits will be ignored, words that contain an `@` (assumed to be an email address) will be ignored. There is an attempt to spot names by their context (viz: ignoring material after a "`Mr`" or "`Mrs`"). Words that are not very interesting (so called "noise words" like `and`/`but`/`the`/`or`/`a` etc) will be discarded.

49. It is intended that postcodes will also be ignored, but this has not yet been implemented.

50. The words are then ranked by frequency, and words that only occur once are discarded. Finally the top 10 in the frequency list (assuming that there are 10) will be retained as a representation of what the web page was all about.

51. All of this work is done by a machine based at the ISP called a "Profiler". This machine of necessity sees all of the web sessions, and it is aware of the IP address of the user whose session is being analysed. It also picks out the UID which uniquely identifies the user from the cookie that accompanied the web page request.

52. Once the analysis is complete then a record consisting of the URL that was visited, the search terms (if any), the top 10 words and the UID is passed to a machine called the "Anonymiser".

53. The Anonymiser passes the record {URL/search/UID/words} across to another machine called the "Channel Server". The Profiler and the Anonymiser are controlled by the ISP, albeit running software supplied by Phorm, but the Channel Server is controlled by Phorm. One instance of the Channel Server function is provided at each of the participating ISPs.

54. The Profiler could have passed the {URL/search/UID/words} record directly to the Channel Server. Phorm say that they pass it via the Anonymiser machine so that it is architecturally obvious that the distillation of the full knowledge of the web session performed by the Profiler has resulted in an "anonymised" blob of data that can be safely processed in more intrusive ways.

55. The Anonymiser is also the machine that provides the cookie processing and the "binding" (multiple `307` redirection) process described in the first section of this document. It also handles the serving of adverts, as will be described in the next section.

56. Anyway, to return to the explanation of the data processing procedure: the Channel Server takes the {URL/search/UID/words} record and processes this against a database to determine all of the "channels" that match.

57. For example, advertisers in the travel business may be looking to promote their wares to someone who was visiting sites containing words like "`flight`" or "`hotel`". If these words appear in the search terms or the frequent word list, then it can be assumed that the web page was about "travel" and hence the advertisers will be interested in advertising to this particular user.

58. Having banged all the rocks together to determine what matches can be made between the distilled information about the web page and the list of words defining the channels, what is finally recorded onto disk are records saying {channel/UID/datestamp}. The information about the URL, search terms and frequent words is immediately discarded.

59. These {channel/UID/datestamp} records are processed when an advertiser specifies something like, "I would like to advertise to people who have visited three or more travel sites in the past week".

60. Necessarily, the advertiser's definition of what is a "travel site" must be available at the point at which the matching against the channel definitions is made. The dynamic part of the specification "three or more in the past week" is only performed when an advert is about to be served.

61. The {channel/UID/datestamp} records are discarded in accordance with the timeout rules that are specified for the particular channel. The maximum period permitted for targeting rules is six months, hence no records will ever be more than six months old.

E Serving advertisements

62. In this section I report how the advert serving system works.

63. Early speculation about the Phorm system suggested that it added adverts to web pages, or replaced them "on the fly". This is not what happens, the specially targeted adverts only appear on participating websites.

64. A website that contains adverts that come from Phorm's "OIX" network will place into their webpages some HTML such as `<img src="http://webwise.net/advert">`, much as they would do today with existing advertising systems. In practice there may be other stuff going on, but in essence it is this simple.

65. The user's browser will therefore fetch this image from `webwise.net` and display it. Along with the request for the image, the browser will automatically send the cookie for `webwise.net` (this is standard behaviour), which will contain the UID by which Phorm knows the user.

66. The request will be routed by the ISP to the Anonymiser machine, which will forward the UID to the Channel Server (which will not learn, for example, the IP address).

67. The Channel Server will then determine which advertisers currently wish to advertise to this UID (given the past history of the types of pages which have been visited and when these visits happened), and will then run a real-time auction between them to decide whose advert will actually be displayed on this occasion. The most valuable advert (to Phorm's bottom line) will then be chosen and the Anonymiser will be told what advert should be served.

68. If the system has no information about a particular UID, then the advert served will clearly not be based on past behaviour by the particular user.

69. This process also involves a second cookie which contains a "frequency cap", whose value is also passed to the Channel Server (to prevent the same advert being shown to the same person all the time).

70. The Channel Server will also be aware of an identifier (a "tag ID") associated with a particular advert position – which may be specific for a particular website or more generically associated with an advertising network. The "publisher" associated with this tag will set rules about what adverts they will accept. This prevents unwanted clashes, for example The Financial Times may wish to ensure that the blank spaces on their pages should never contain adverts for The Daily Mail ...

71. The Channel Server exists within each ISP. Conceptually it could have been a central resource, but subsequent to the meeting they have explained that the reason for providing this functionality locally is performance and availability reasons.

72. In the meeting itself, Phorm said that the reason for not centralising the Channel Server function was that they were concerned about EU Regulations concerning moving personal data outside of Europe.

73. I then pointed out that it was Phorm's contention that the {channel/UID/datestamp} information was completely anonymised and hence there would be no movement of data that could fall foul of the regulations – to which they responded that they wished to ensure that they were not being perceived to break the rules, even though they did not believe they would do so.

74. The Phorm design includes provision for "roaming" so that if your UID is used on one participating ISP at work, and another participating ISP at home, then the data could be linked together. Note that to implement this would involve more than just the Channel Server functions of the different ISPs communicating with each other.

75. Roaming is enabled by default within an individual ISP, viz: where the user connects to different "POPs" (points of presence) at different times. Roaming is not enabled by default for roaming between ISPs. Phorm believe that EU Data Protection regulations mean that they could not enable roaming to ISPs outside of Europe.

76. Clearly it would be possible to create a channel containing unusual words so that only users who visited particular web pages would match that channel. Phorm say that they address this type of issue firstly by vetting the words that are used to define a channel, and secondly by insisting that a channel contains at least 5000 UIDs before any adverts are served that are related to it.

77. It is Phorm's belief that it is not possible for the Channel Server (viz: the part of the system operated by them) to make a link back to any particular individual.

78. With detailed knowledge of which sites were visited by a particular user, it might be possible to inspect the channel records and determine their UID. Conversely, given their UID, the channel records are unlikely to reveal very precise information about their activities. Phorm do not believe that either scenario is realistic.

79. Although linkage (to IP addresses if not to individuals per se) can be performed by inspecting data held by the Profiler or the Anonymiser machines, this does not make any theoretical difference to what the ISP (who operate this equipment) is able learn about their customers. Furthermore, no obvious practical reason for inspecting this data is apparent to Phorm – the systems are, for example, completely inappropriate for use in warranted interception.

80. The Phorm system will not be creating channels which relate to advertisements for adult material, for anything medical, or for alcohol, tobacco, gambling or politics. This means that they will not be matching words with these themes, or serving adverts for this type of material.

F Anti-phishing

81. The Phorm system provides an anti-phishing mechanism. They are buying in lists of phishing sites, currently from two well-known commercial companies.

82. If a user attempts to visit a page which is on this list then the Layer 7 switch will direct them to a machine which will serve up a warning page. If the user chooses to continue to website regardless, a cookie will be used to record this choice so that the warning page does not keep on cropping up. The individual ISPs will specify how long this cookie will last for.

83. Existing documentation says that the anti-phishing mechanism will not be available if the user has opted-out of the Phorm tracking mechanisms. This description suggests that there is no technical reason for this; provided that the opt-out isn't applied at the ISP level so that the Phorm equipment is completely bypassed.

84. Phorm were vague as to the extent to which they would be using regular expressions to detect phishing URLs. At present, handling the URLs for the rock-phish gang, and some other high-profile groups, is ineffective unless generic matching techniques are used.

G Other matters

85. Phorm have previously reported on a (favourable to their offering) user survey which was conducted by their partner ISPs. This was conducted by a reputable market research company, who surveyed 2500 people and conducted extensive focus groups.

86. In the course of the meeting Phorm made a number of statements and claims about their system. I include them in this section for completeness. Please note that these are Phorm's opinions, not mine.

    (a) Phorm state that it is impossible for them to know who they are serving adverts to.

    (b) Phorm state that it is impossible for them to know exactly where you have been and what you have done.

    (c) Phorm believe that their system represents a true choice for consumers; whose data is currently processed in a less privacy-preserving manner.

    (d) Phorm argue that their system should be viewed as a PET (a "Privacy Enhancing Technology") and that it should be welcomed by Data Protection Registrars as demonstrator of commercially viable technology that makes it unnecessary to collect large amounts of identifiable personal data, the way in which existing systems operate.

    (e) Phorm say that their conversations with the Office of the Information Commissioner have gone "extremely well". Phorm believe that they have "gone to great lengths to protect data and enhance privacy".

(f) Phorm said that, in their view, the key to compliance of their system is the nature of the words that are looked for within their channels. Words and URLs for new channels are checked against a "blacklist" of forbidden items and against the "whitelist" of existing "vetted" channels. Any new terms are manually inspected.

(g) Aside from the inability to link UIDs to people, the lack of a medical channel means that, in Phorm's opinion, the scenario recently envisaged by Sir Tim Berners-Lee (that an medical insurance company could raise rates because he was reading informational websites about cancer) could not occur.

(h) Phorm and their partner ISPs approached the Home Office to seek a letter of comfort that their system did not perform illegal "interception", within the meaning of the Regulation of Investigatory Powers Act 2000. The document that resulted was recently published by the Home Office.

(i) Phorm have also obtained a QC's opinion on whether they are performing illegal interception. I asked if anyone had ever told them their systems would be illegal to operate and if they had subsequently modified them, and was told that, "we didn't get any opinions that it was illegal".

(j) Phorm observes that their system turns round the situation for website owners (who wish to serve advertisements). The adverts served from a particular page will not depend on the content of the page (the current model, supported by Google &c), but will depend on the interests of your visitors.

(k) Phorm believes that their system will increase revenue to websites that can build an audience, leading to better "free" content on the web. Phorm thinks that this is a good thing for the Internet as a whole.

(l) Phorm wish to reach a consensus, through open debate, as to how "behavioural advertising" should be operated going forward.

Dr Richard Clayton
Cambridge, UK
4$^{\text{th}}$ April 2008