

Конфиденциальность информации в Интернете

Андрей Робачевский
Технический директор RIPE NCC

Десятилетие лет назад совершение онлайн платежа считалось делом рискованным и делалось с большой осторожностью. Использование реальных имен в социальном общении в Сети было немислимо, а анонимность деятельности в Интернете считалась почти абсолютной.

Социальные сети радикально изменили ситуацию. В социальных сетях мы хотим общаться не с незнакомцами или виртуальными персонажами, а с людьми, которых мы знаем или хотим познакомиться. Это предполагает, что наш образ в сети достаточно правдив.

Мы обмениваемся со своими друзьями, знакомыми, а зачастую и со всем миром, вполне правдивой информацией о своем возрасте, интересах, местонахождении, текущими заботами, да практически обо всем!

Например, недавно я прочитал в газете о сайте Blippy (blippy.com), позволяющим обмениваться информацией о сделанных покупках. Некий Марк оповещает весь мир о том, что он приобрел кейс для iPad'a за \$41, потратил \$24 в ресторане Applebee и \$6450 в клинике пластической хирургии во Флориде на операцию носа. Уникальный случай? Похоже, что нет. Сайты и стоящие за ними социальные сети, для обмена самой разнообразной каждодневной информацией растут как грибы после дождя и пользуются огромной популярностью. Например Foursquare – мобильная социальная сеть, позволяющая анонсировать ваше местонахождение. Или SpurtUp, откуда мир может узнать, сколько отжиманий вы сделали сегодня.

На первый взгляд, публикация в Интернете пусть даже персональной, но довольно безобидной информации, например о любимых занятиях, собственной фотографии или номера школы, которую вы закончили, не представляет опасности. Однако сайт PleaseRobMe.com наглядно показал, как информация социальных сетей может быть использована в совсем не безобидных целях. Этот сайт использовал информацию доступную с сетей Twitter и Foursquare для обнаружения пустующих квартир, хозяева которых проводили время вдали от дома.

Другим аспектом является то, что область распространения вашей личной информации может оказаться гораздо больше, чем вы думаете. Она зачастую не ограничена узким кругом друзей и знакомых, и, более того, почти всегда выходит за рамки вашей социальной сети. Это, в свою очередь, может существенно ограничить ваши возможности оставаться анонимным в случаях, когда вы этого хотите.

В 1993 году в журнале New Yorker была опубликована карикатура, изображающая двух собак, сидящих перед дисплеем компьютера, одна из которых говорит другой: В Интернете никто не знает, что ты собака (<http://www.newyorkerstore.com/Dogs/On-the-Internet-nobody-knows-youre-a-dog/invnt/106197>). Сегодняшняя реальность существенно отличается от этой картины.

Давайте рассмотрим эту проблему более подробно.

Утечка и Консолидация информации

Начнем с вопроса, насколько конфиденциально пользование Интернетом вообще? На первый взгляд, использование Интернета весьма анонимно. Конечно веб-сайт, который вы посетили, знает IP-адрес вашего компьютера, но что из того? отдельный сайт и обезличенный адрес. Вроде бы нет повода для беспокойства.

Однако многие, если не большинство посещаемых веб-сайтов используют те или иные технологии для отслеживания посетителей. Даже если это и не портал с пользовательским именем и паролем, всякий раз, когда вы заходите на сайт (а сегодня это почти синоним работы в Интернете) происходит "утечка" частной информации о посетителе сайта, то есть о вас. Например, популярными технологиями являются встроенные "жучки" в виде изображений в 1 пиксель, cookies или приложения JavaScript.

Утечка частной информации также происходит не только на самом веб-сайте, который вы посетили, но также и к сайтам третьих сторон. Эти сайты часто, хотя и неочевидно, присутствуют на просматриваемой веб-странице, например, в виде рекламных фрагментов. Типичным сценарием является размещение рекламных объявлений рекламными провайдерами, например Google's AdSense (googlesyndication.com, doubleclick.net), AOL (advertising.com, tacoda.net), Yahoo! (yieldmanager.net), по договоренности с владельцем сайта. Обычно эти фрагменты существуют в виде компонентов JavaScript или просто графической картинке. Отображение страницы современного информационного веб-сайта

включает десяток обращений за различными элементами, в том числе и к сайтам третьих сторон. Стоит отметить, что это происходит независимо, кликнул ли пользователь на банер или нет.

Другим вариантом третьих сторон, присутствующих на сайте являются компании, предлагающие аналитические услуги (посещаемость, трафик, клиенты, тенденции), такие как например google-analytics.com (Google), 2o7.net (Omniure), atdmt.com (Microsoft/aquantive), quantserve.com (Quantcast).

Наконец, сети распределения контента (CDN), akamai.net, yimg.com и т.п., размещают изображения и видео.

Все это ручейки, по которым информация о вашем визите распространяется гораздо дальше, чем сайт, который вы посетили.

На пленарном заседании IETF78 исследователь Balachander Krishnamurthy из AT&T Labs-Research сделал интересный доклад, посвященный теме утечки личной информации в Интернете (<http://www.ietf.org/proceedings/77/slides/plenaryt-5.pdf>). Эта презентация основана на многолетнем исследовании, проведенном совместно с Craig E. Wills из Worcester Polytechnic Institute. Ученые изучали так называемое "пятно конфиденциальности", определяющее степень распространения пользовательской информации среди с виду несвязанных сайтов. Исследование состояло из 9 экспериментов, охватывающих период в 5 лет, 1200 наиболее популярных сайтов (по данным Alexa, <http://www.alexa.com/>) в различных категориях, 68 странах и 19 языках.

Одним из исследуемых параметров являлась так называемая степень ассоциации посещаемых сайтов друг с другом через сайты третьих сторон. Например, если два независимых сайта используют google-analytics.com, они считаются ассоциированными. Результаты получились довольно любопытными. 70% всех исследованных посещаемых сайтов имеют ассоциацию с более 400 несвязанными сайтами.

Но еще поразительнее оказалась степень концентрации сайтов третьих сторон. Всего 10 крупнейших узлов третьих сторон (doubleclick.net, google-analytics.com, 2mdn.net, quantserve.com, scorecardresearch.com, atdmt.com, omniure.com, googlesyndication.com, yieldmanager.com, 2o7.net) оказались представленными на 78,5% всех посещаемых сайтов.

Но и это еще не все. Анализ покупки и слияний компаний в этом секторе рынка позволил исследователям объединить большинство этих узлов в три крупнейших семейства: - Google, куда входят, например, doubleclick.net, googlesyndication.com и google-analytics.com - Adobe, где крупнейшими представителями являются omniure.com, offermatica.com и hitbox.com, и - Microsoft, с его крупнейшим приобретением - Aquantive - в 2007 году.

Менее значительными, но также заметными являются "семейства" Yahoo и AOL.

Наблюдение за распространением этих семейств на посещаемых сайтах показало его постоянный рост, с 40% в 2005 году до 84% в марте 2010!

Глубина проникновения также увеличилась. Это означает, что посещения пользователя отслеживают более одного семейства сайтов третьих сторон.

Таким образом в распоряжении компаний, предоставляющих эти услуги, находится громадный объем информации о работе пользователей в Интернете, а возможность корреляции данных о посещениях независимых сайтов позволяет создать "профиль" пользователя, определить его вкусы, привычки, взгляды, местоположение и другие индивидуальные характеристики.

Правда, когда мы говорим "пользователь", речь в большинстве случаев идет о безымянном IP-адресе компьютера, или компьютеров, на которых работает пользователь. Другими словами, можно получить довольно надежную информацию о некотором пользователе, но без его идентификации.

Утечка конфиденциальной информации

Этот аргумент часто используется компаниями, предоставляющими услуги сайтов третьих сторон. Однако Balachander Krishnamurthy и Craig E. Wills исследовали вопрос более детально и показали, что "пользователя" можно идентифицировать, благодаря социальным сетям. Более подробно с этой работой можно ознакомиться на сайте <http://www.research.att.com/~bala/papers/pmob.pdf>.

Не секрет, что большинство сайтов социальных сетей используют услуги третьих сторон, о которых я только что рассказал. Это и размещение рекламных объявлений, контента, а также разнообразных счетчиков и треккеров, позволяющих отслеживать перемещение пользователя в веб-пространстве.

Хуже, когда обращение к сайтам третьих сторон содержит идентификатор пользователя социальной сети. Например, отображение домашней страницы vkontakte.ru включает обращение к сайту counter.yadro.ru:

```
GET
/hit?rhttp%3A//login.vk.com/%3Fact%3Dlogin;s1920*1200*24;uhttp%3A//vkontakte.ru/id12...7;0.134937964353806
HTTP/1.1
Host: counter.yadro.ru
Referer: http://vkontakte.ru/id12XXXXX7
Cookie: VID=1YnD5w3WMDml
```

Как видно, идентификатор пользователя (**id12xxxxx7**) содержится в этом обращении.

Конечно, объем информации, который можно получить о пользователе, зная его идентификатор, зависит от конкретной сети и настроек пользователя, но анализ наиболее популярных сетей все теми же американскими исследователями показал что имя, личная фотография, пол, увлечения, список друзей, образование в большинстве случаев открыты по умолчанию.

Теперь третья сторона может ассоциировать различные атрибуты пользователя, например IP-адрес его компьютера, с конкретным персонажем социальной сети. Подобная ассоциация может быть установлена и с cookie-треккером, также часто передаваемая в запросе. Поскольку cookie-треккеры как правило являются долгожителями (например, срок действия cookie в приведенном примере истекает в апреле следующего года), прошлые и будущие "путешествия" пользователя могут быть ассоциированы с этим персонажем и таки образом идентифицированы.

Принимая во внимание степень ассоциации сайтов и концентрации сайтов третьих сторон, возможности третьих сторон по отслеживанию и "профилированию" деятельности пользователей в интернете значительны. Конечно не факт, что третьи стороны производят или используют подобные ассоциации, но потенциальная возможность для этого имеется.

Учитывая, что большинство сайтов третьих сторон принадлежат компаниям так или иначе связанными с рекламным бизнесом, эти возможности вряд ли используются в политических или криминальных целях. Сегодняшний «Большой Брат» активно зарабатывает деньги, предлагая нам все более утонченную рекламу и, как никогда, управляет нашим выбором.

Зачем нужна конфиденциальность

Даже отбросив экстремальные случаи, один из которых я упомянул в начале статьи, большинство из нас не видит большой опасности от публикации собственной фотографии, списка друзей или возраста. Или названия города, в котором я живу, моих любимых авторов и фильмов.

В конце концов, даже информация о популярных сайтах, которые я посещаю, не такой большой секрет. Другими словами – мне особенно нечего скрывать.

Возможно это и так. Однако проблема в том, что хотя мы можем оценить ущерб от публикации отдельного фрагмента данных о себе, например имени и фото, масштабная агрегация и корреляция многочисленных фрагментов, включая наши предпочтения, перемещения, общения и т.п., может представлять информацию, которую мы предпочли бы хранить при себе.

Возвращаясь к социальным сетям, незнание пользователя о степени публичности тех или иных персональных данных, является объектом серьезной критики. Под давлением этой критики Facebook, крупнейшая из социальных сетей, объявила в конце мая этого года о серьезных изменениях в системе контроля персональных данных. В частности, пользователи получают полную информацию о том как, кому и какие данные являются доступны, а также возможность задать мастер-установки конфиденциальности, применимые ко всем настоящим и будущим приложениям Facebook.

Другой проблемой является то, что пользователь не имеет понятия, о составе, масштабе и использовании этой информации, и, соответственно не имеет возможности оценить последствия. Пользователь также не может исправить возможные ошибки в этих данных.

Немного напоминает Кафку, не правда ли?

Вместо заключения

Возможности обмена информацией в Интернете сегодня поистине безграничны и продолжают стремительно развиваться. Социальные сети создали невиданную доселе степень информационной связности между людьми в глобальном масштабе. Интернет превратился в динамичную социальную среду, объединяющую сотни тысяч миллионов людей. Как сказал основатель Facebook Mark Zuckerberg, миссией компании является сделать мир более открытым и связанным.

Но конфиденциальность является важным элементом открытого информационного пространства Интернета. Пренебрежение правом человека на конфиденциальность, а точнее его правом контроля за составом и степенью распространения персональной информации, приводит к потере пользовательского доверия, усилению контроля и как следствие к уменьшению инновативности, интенсивности и объема информационного обмена.

Поэтому очень важно поддерживать этот сложный баланс между конфиденциальностью и открытостью, удобством пользования и утечкой частной информации. В какой-то степени об этом могут позаботиться сами пользователи, через установки браузера или социальной сети. Например, блокирование заголовка Referer в запросе или запрет приема cookie от сайтов третьих сторон. Однако эти меры часто малоэффективны. Например, отказаться от использования cookie посещаемого сайта сегодня практически невозможно без существенной потери функциональности.

Существенный вклад в решение этой проблемы могут внести создатели самих сайтов и социальных сетей. Например, минимизируя количество информации, передаваемой третьим сторонам, или информируя пользователей о степени открытости их данных в социальной сети. И делать это не дожидаясь судебного процесса, как это произошло с услугой Facebook под названием Beacon, на которую были автоматически подписаны все пользователи сети в ноябре 2007 года. Благодаря Beacon друзья пользователей получали оповещения о деятельности последних на некоторых других сайтах, например, о покупке билетов в кино на сайте Fandango (<http://www.fandango.com/>). Дело закончилось двумя годами спустя полным прекращением услуги и созданием Facebook фонда для работ в области онлайн конфиденциальности размером в 9,5 миллионов долларов.

Знание и открытое обсуждение этих проблем – уже шаг к их решению. И проблемы действительно решаются: пользователи получают больший контроль в социальных сетях, браузеры заботятся о нашей безопасности и конфиденциальности. Потому что несмотря ни на что, люди хотят обмениваться информацией со своими друзьями, знакомыми, а иногда и со всем миром.

Что ж, пора обновить фотографию в Одноклассниках ...

Андрей Робачевский, Технический директор RIPE NCC

Мнения, представленные в статье, не обязательно отражают официальную позицию RIPE NCC