Limiting Large-scale Crawls of Social Networking Sites

Mainack Mondal
MPI-SWS
mainack@mpi-sws.org

Bimal Viswanath
MPI-SWS
bviswana@mpi-sws.org

Allen Clement
MPI-SWS
aclement@mpi-sws.org

Peter Druschel
MPI-SWS
druschel@mpi-sws.org

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Ansley Post
MPI-SWS
abpost@mpi-sws.org

ABSTRACT

Online social networking sites (OSNs) like Facebook and Orkut contain personal data of millions of users. Many OSNs view this data as a valuable asset that is at the core of their business model. Both OSN users and OSNs have strong incentives to restrict large scale crawls of this data. OSN users want to protect their privacy and OSNs their business interest. Traditional defenses against crawlers involve rate-limiting browsing activity per user account. These defense schemes, however, are vulnerable to Sybil attacks, where a crawler creates a large number of fake user accounts. In this paper, we propose Genie, a system that can be deployed by OSN operators to defend against Sybil crawlers. Genie is based on a simple yet powerful insight: *the social network itself can be leveraged to defend against Sybil crawlers.* We first present Genie's design and then discuss how Genie can limit crawlers while allowing browsing of user profiles by normal users.

General Terms

Security, Design, Algorithms

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*

Keywords

Sybil attacks, social networks, network-based Sybil defense

1. INTRODUCTION

Online social networking sites (OSNs), such as Facebook, Twitter, and Orkut, contain data about millions of users. These OSNs allow users to browse the profile of other users in the network, making it easy for users to connect, communicate and share content. This core functionality of OSNs, however, can be exploited by crawlers to aggregate data about large numbers of OSN users for re-publication [1] or other more nefarious purposes [2] that violate users' privacy.

SIGCOMM'11, August 15–19, 2011, Toronto, Ontario, Canada.
ACM 978-1-4503-0797-0/11/08.

Crawlers present a significant problem not only for OSN users but also for OSN site operators. First, many OSNs view the user data as a valuable asset that could be leveraged to generate revenue in the future, for example, via targeted advertisements. So OSNs have an incentive to prevent third party crawlers from accessing their data. Second, while OSN operators can ensure that data is used according to privacy policies specified on their sites, they cannot make any guarantees about how crawlers will use that data. A third party that crawls an OSN can do anything with that data (e.g. republish the data or infer private information [2]). Yet, if the third party crawler does something nefarious, the OSN operator is likely to be held responsible, at least in the court of public opinion. For example, Facebook was widely blamed in the popular press for allowing a crawler to gather public profiles of a large number of users [1].

Today OSN operators employ various rate-limiting techniques to restrict a crawler's ability to scrape the network. These techniques typically rely on limiting the number of user profiles a single user account or IP address can view in a given period of time [4]. Unfortunately, these schemes can be easily circumvented by a Sybil attack, in which the crawler creates a large number of fake user accounts and/or hires a botnet to gain access to multiple IP addresses.

In this work, we propose Genie, a system that OSN operators can deploy to limit Sybil crawlers. Genie relies on a key assumption about OSNs, namely, that *it would be hard for crawlers to establish an arbitrarily large number of links to users in an OSN.* Intuitively, this assumption is based on the observation that forming a new link between two users requires a certain amount of familiarity between the users involved. Genie leverages this insight to limit large scale crawls. It ties the ability of a user to crawl the OSN to the number of links she establishes with the rest of the network.

2. GENIE DESIGN

At its core, Genie models the trust between nodes in a social network as a credit network [3] and leverages this network to ensure crawlers cannot collect additional information by creating many identities. Specifically, Genie maps the nodes and links in the OSN to nodes and edges in a credit network. Each edge in the credit network is assigned some initial credit that is refreshed periodically. User A is allowed to view the profile of user B only if there exists a path in the credit network that connects the two users and has sufficient

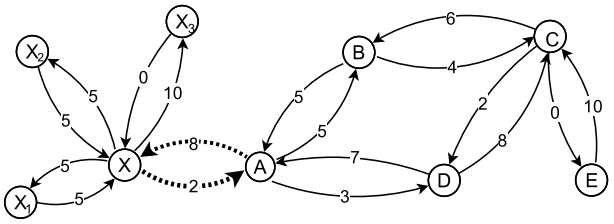

Figure 1: Genie represents the social network as a directed graph with available credit on the links as shown above. X is a crawler with Sybil identities (X_1, X_2, X_3). The links between crawlers and the rest of the network are shown with bold dotted lines.

credit on each link along the path. If no such path exists, then user A is blocked from browsing the data of user B.

Figure 1 illustrates how Genie represents the social network as a credit network consisting of directed links and leverages the credit network to determine which profile visits are allowed. For example, in the figure, Genie allows user A to visit the profile for user C because there exists a directed path from A to C with every edge on the path having a credit value of at least one. Once A visits C's profile, the credit on each directed edge in the selected path is decremented by one. User A, on the other hand, is not allowed to visit the profile for user E because there is no path from A to E in which all edges have sufficient credit to support the profile visit.

In the above example, we focussed on a pricing model that deducts one unit of credit from each edge along the path from A to C. We note, however, that there exist different pricing models that bound the ability of users to browse other users' profiles to different extents. While the pricing model that best suits an OSN depends on the OSN's workload (i.e., users' browsing behavior) and the OSN's social graph, Genie ensures that the following property holds for any valid pricing scheme: *Given any cut through the social network graph, the rate of profile view requests between nodes on different sides of the cut is proportional to the number of edges in the cut.* Thus with Genie, the ability of a user to crawl the OSN is tied to and limited by the number of social links she establishes with the rest of the network. Coupled with the observation that it is hard for crawlers to establish arbitrarily large number of links to other OSN users, this suggests that, even when a crawler creates a large number of Sybil identities, her ability to crawl the network would be limited.

We illustrate this situation in Figure 1 where crawler X creates three Sybil identities: X_1, X_2 and X_3. However, these Sybil identities do not give the crawler any extra benefit since the available credit on the cut between the crawler nodes and the rest of the network remains the same. Thus the crawlers, having a limited number of links to the rest of the network, would be allowed to perform only a limited number of crawls.

While the basic operation of Genie is straightforward, there are a few obvious points of concern:

Does Genie restrict access to popular content? A side effect of deploying Genie is that it limits the total number of profile views (exposures) that a OSN user may receive to the amount of credit on all the incoming links to the user. We argue that such a restriction is often in the interests of ordinary users, many of whom make their personal data accessible to the public, under the implicit (though incorrect) assumption that their data would not be viewed by the whole world. There is a subtle but important difference between data being *accessible* to the public at large and data being *viewed* by the public at large. Thus, we argue that the limit Genie imposes on the number of views a user's data might receive, even as the data is accessible to the public at large, might be a desirable side-effect.

On the other hand, certain users, such as politicians, celebrities, and marketeers, might not want the profile views that they receive to be limited by Genie. We would argue that these users would in any case have many friend links, which would naturally allow more views. A side-effect of this approach is that a celebrity would be able to crawl a large number of users. But it is hard to imagine a celebrity behaving like a Sybil crawler. Some of these celebrities may further desire that their profile could be viewed by everyone without any limits. OSN sites can accommodate such users by explicitly allowing them to keep some or all of their profile contents outside the Genie framework. Access to the content would not be moderated by Genie and the content would not be protected from aggregation by crawlers.

Does Genie introduce new DDoS attacks? If Genie is deployed it may be feasible for a group of OSN users to launch a DDoS attack against a specific user A or against a small group of users. They can visit user A's profile repeatedly, exhaust credit on A's incoming links, and thus, block further access to A's profile for everybody, including A's friends. Genie solves this problem by allowing user A to create a whitelist of users (e.g., all direct friends of A), who would be granted access to A's profile regardless of the availability of credit. This ensures that the user A's profile data is always available to her direct friends.

3. DEPLOYMENT CHALLENGES

The primary challenge that OSN operators will face when deploying Genie lies in managing liquidity (i.e., credits) in the network. The OSN must decide upon the pricing model, the amount of credit value that is initially assigned to the directed edges of the OSN, and the rate at which credit is refreshed. The credits should be managed such that it would allow normal user browsing activities, while limiting the crawler or data aggregator activities. If an OSN assigns too much credit, it would allow too much of the undesired crawling activity. But, if it assigns too little then it could affect normal browsing activities of users.

We are currently investigating methods to determine suitable credit values by analyzing past user behavior. We are experimenting with different settings and have some promising preliminary results. We are still in the process of completing a thorough evaluation of the system. For more details, please refer to `www.mpi-sws.org/~mainack/genie/`

4. REFERENCES

[1] `http://tcrn.ch/9JvvmU`.
[2] `http://bit.ly/jlarLI`.
[3] D. DeFigueiredo and E. T. Barr. Trustdavis: A non-exploitable online reputation system. In *CEC'05*.
[4] T. Stein, E. Chen, and K. Mangla. Facebook Immune System. In *SNS'11*.