Social Networking Sites:
Data Mining and Investigative Techniques

Stephen Patton, CISSP

August 2007

Social Networking Sites

It is hard to deny the booming popularity of social networking sites, the type of sites that facilitate a high degree of user personalization, and user intercommunication. While yearly growth in the largest sites may have started to slow, there is evidence that growth is accelerating in communities that have previously not had a high degree of social networking site use.

185 Million MySpace Users (April, 2007)

News articles have cited social networking site use rates as high as 60% for teenagers aged 13 to 17. Principals and teachers are using social networking sites to better understand, and sometimes to monitor activities of their students. Local law enforcement is using these sites to solve crimes. In some cases, law enforcement has been able to use social networking sites to prevent serious crimes before they occurred.

The unnerving level of candid content posted by teenagers has led to a number of cases involving stalking, abduction, and worse. Many of these sites are fertile ground for social engineering due to the level of disclosure routinely engaged in by users of the sites. Law enforcement in many areas is involved in many related activities including the investigation of crimes tied to social networking sites and ongoing education programs to sensitize parents and children to the risks of over disclosure in cyberspace.

Wired Magazine journalist Kevin Poulsen data mined MySpace in 2006 to discover hundreds of sex predators registered in their own names with active profiles on the site. His technique involved matching Department of Justice registered sex offenders with profiles on the site with automated scripts. Since that time MySpace has made efforts to identify such users and remove them when they can be identified.

40+ Billion MySpace Page Views Monthly

For the investigator, the broad use of these sites raises the likelihood of finding useful information on one or more of these sites. Even less popular sites than MySpace such as Facebook, Bebo, and many others each have tens of millions of users. The possibility of various evidence gathering activities is quite promising with the combination of high use ratios and high disclosure use patterns on the part of social networking site users.

Data Mining Heaven or Hell?

With all the hype surrounding social networking sites, it would seem this area is the perfect environment for extensive data mining, research, and development. While the usage rates, public availability and media scrutiny all point to increased interest in the sites, there are a number of impediments to capitalizing on data mining strategies for this area.

From an academic research perspective, studying contact networks, growth rates, and social implications of social networking sites is not likely to draw strong opposition from site owners or users. In fact, sites may be supportive of these activities if they are perceived as supportive of the sites and the environment they try to foster.

350,000 new MySpace registrations daily

It has already been shown that public safety research on such sites is welcome and supported by the public and site owners. Anything that is perceived as addressing a safety issue for children is not likely to be strongly opposed by site owners or users in public. To the extent that law enforcement interest in social networking sites is viewed as a child safety initiative, there will likely be a reservoir of public good will towards such activities, in spite of the obvious possibility of abuse from over collection of information or broad mining activities in pursuit of other law enforcement goals.

Along with such challenges as terms of service limitations, potential copyright issues, and significant privacy issues, there are many technical issues that will diminish the effectiveness of any broad data mining initiative with social networking sites. The arena of social networking sites is highly fragmented. While MySpace enjoys unchallenged dominance in the category, there are many sites catering to specific needs: LinkedIn for working professionals, Classmates.com for reunions of high school and college classmates, Facebook for students, and Friendster, Bebo and others for similar activites. Each of these sites is unique. Each has a specific structure for linking information to user profiles, and each would need custom clients for data collection.

4.5 million concurrent MySpace users

As if this fragmentation were not enough, changes to the site can invalidate any data collection client instantly, rendering the site unable to be mined until the client is updated. Additionally, information retrieved will not likely be compatible with other sites, and there are no standard ways to link information across social networking sites, rendering each a separate island from all the others. This should not be a surprise given the lengths site owners go to in developing and maintaining brand and user loyalty. Not only can changes by the site owner cause collection problems, but a variety of sites permit varying degrees of customization by users that can further contribute to collection problems.

There is also a shift toward non-text content in the form of sound, music, pictures and video that render basic data mining techniques ineffective in recovering content that can be interpreted and collated by means of automated software. Attempts to recover and interpret such content would be technically challenging and fraught with legal danger from illegal content of various types.

Media scrutiny of social networking sites is pushing many users to designate their profile private. This makes the profile inaccessible to those not granted permission from the profile owner. While this certainly improves user privacy, it is a growing challenge to data mining as it removes much content

from the sphere of recoverable data. MySpace profiles are generally public unless designated private, whereas profiles on Facebook are only visible to those designated friends by the profile owner. Such differences create a dramatically different picture of what data is available on a particular site.

1 Billion images and 25 Million songs on MySpace

A remaining challenge to text based content recovery and interpretation is essentially semantic in nature. Given the level of slang, abbreviations, and assumed knowledge on the part of social networking site participants, interpreting recovered conversations and content can be quite difficult.

Yet there is a fair amount of useful information which can be recovered, and past experience has shown that intelligent interpretation of profile content can lead to crime prevention or detection, greater public and/or child safety, and other useful results.

Site Structure and Approach

Regardless of the target site, content recovery should focus on text. Text is the most readily usable content on user profiles, and the least likely to threaten the researcher with a variety of content based risks. These risks primarily include copyright issues for multimedia content and sex offender charges for possession of child pornography.

Due to the leadership of MySpace in the social networking site market, the analysis and tools for this research focuses on MySpace. That has lead to a number of design and implementation decisions. In order to simplify the tool set and demonstrate the relative ease by which content recovery can occur, the tools are written exclusively in Perl. Additionally, more complex Perl packages could certainly have been employed, such as tidy and XML parsing. One reason this course was not followed is that the condition and quality of HTML compliance on MySpace is quite low. Since context-based cues were effective in quickly finding the relevant content, the “more robust” approach was not taken. The core tools are under 300 lines of Perl, and the entire code submission is under 500 lines of Perl.

Toolz

The code accompanying this paper includes important information and warnings about appropriate use of these tools. Any use of these tools is completely at the risk of the user. All scripts were tested and in working order at the time of release. The author is not able to provide support for these tools.

The six short Perl programs will be introduced here. For brevity, the code will not be duplicated in this document.

myspaceminer_login.pl: The login script requires no arguments. It must be modified by the user to provide an ID and password for MySpace access. This is required to successfully recover content past the initial profile page. The script makes a cookies file that is used by myspaceminer_scrape.pl.

myspaceminer_scrape.pl: This script takes a profile name (or friendid number) as an argument. It recovers the profile page, and comment pages. These are stored in plain text in the local directory.

myspaceminer_comment.pl: This script takes a profile name (or friendid number) as an argument. It processes all the comment pages of the profile, and produces delimited lines of comments suitable for loading in a database.

profile_comments.pl: This script takes a profile name (or friendid number) as an argument. It produces delimited lines of comments from the profile page (only) that are suitable for loading in a database.

profile_links.pl: This script takes a profile name (or friendid number) as an argument. It produces a list of HTTP links found on the profile page.

comment_converse.pl: This script takes two profile names (or friendid numbers) as arguments. It produces a time ordered list of delimited lines of comments. The comments are either to or from one of the profile names.

Investigation

To review recoverable content of a profile, execute the following scripts:

1. myspaceminer_login.pl
2. myspaceminer_scrape.pl <profile name>
3. myspaceminer_comment.pl <profile name> > comment.file

The file comment.file should now contain the text of all comments of the profile. Keep in mind that many comments contain graphics as part of the message, or pictures with embedded text, and this content is not recovered. TIP: If the comments are not recovered, try running myspaceminer_comment.pl with the verbose flag and review the HTML triggers while comparing to the raw HTML saved in the local directory.

Run profile_links.pl on the profile to determine if any interesting photo albums or other references are linked by the profile. TIP: depending on the photo album site and the album owner's preferences, you may find valuable publicly available photo content by manipulating the URL of a photo URL in the links list.

Successive use of scrape on related profiles (those with comments on the first profile, or in the friends list) will start to produce a related body of recovered content for further investigation. An easy addition would be to recover the friends list of a profile. Since comments reflect “active friends,” this tool set favors comment analysis over friend list analysis.

Just a Start

These tools are just a start at a more comprehensive tool set. They do not recover all available content on MySpace, and successive sets of tools would be required for other social networking sites due to unique site structure for each one.

Of more interest is the analysis of recovered content. In working with local law enforcement, this author has found significant interest in providing relevant analysis for purposes of petty crime investigation. There also may be opportunities to provide prevention intelligence across a range of themes or crime categories, but the challenges of adequately interpreting content are formidable. There are of course significant privacy and public safety issues involved with the use of recovered content, and no doubt one person's public safety initiative will be another's Big Brother initiative. The possibility of using such information for both good and ill is readily apparent. There does seem to be anecdotal evidence that profiles and photo content are “going private” thus excluding them from recovery. Users seem to be responding to media scrutiny by protecting their own content, and though this hampers the work of a researcher, it can hardly be opposed. It seems inevitable that further work will be done in this area, and that both benefits and abuses will be seen in the future.