DSicovery

eDiscovery About People™

WHITE PAPER

The Wild West of Social Media Evidence Collection

www.DSicovery.com
877-797-4771

TABLE OF CONTENTS

INTRODUCTION

When it comes to social media and webmail in forensic collection...these two are a lawless, wild west bunch.

Everyone – from kids to great-grandparents – seems to be using social media and webmail on a daily basis.

Take a look at these statistics:

- 1 in 5 minutes online is spent on social networks[1]
- 6.6 hours a month are spent per user on Facebook[2]
- 400 million tweets are sent everyday[3]
- 72 hours of video are uploaded to YouTube every minute[4]
- 41.5 million new blog posts are published on WordPress every month[5]
- 4.5 million photos are uploaded to Flickr every day[6]
- There are:
  - 3.3 billion email accounts[7]
  - 2.7 billion social networking accounts worldwide[7]
  - 1 billion+ active users on Facebook[8]
  - 200 million active users on Twitter[9]
  - 200 million members on LinkedIn[10]
  - 105 million blogs on Tumblr[11]
  - 64 million WordPress sites[12]
  - 48 million Pinterest users[13]

With that much data being created online, it only makes sense that some of it could be essential to a lawsuit and/or an investigation. Yet collecting the information while maintaining data integrity and review ability is still an untamed land.

Questions arise such as:

- How do the various webmail and email formats become standardized and able to be deduplicated?
- What authorization do you need from service providers to collect information without violating the user agreements?
- What can legally be collected from social media accounts about a user's friends and connections?

LIFE ON THE RANGE

The industry of digital forensics and electronic discovery is still a rather young one. Yet it has been around long enough to develop standards and best practices for handling multiple types of digital files on various mediums.

The data collection process has traditionally been about documents, emails and graphics found on computers, hard drives, phones and other mediums. Now, it also includes data from social networking sites, which requires careful attention and adaptability to ensure the digital information maintains its initial context and meaning.

This challenge of taming the land of social media and webmail – where each platform has its own rules, or no rules at all – is just like taming the wild west. Data collection must be done in a way to fully preserve the information, even if dealing with multiple outside parties and systems for just one social media platform.

Data from social networking sites requires careful attention and adaptability to ensure the digital information maintains its initial context and meaning.

WRANGLING WEBMAIL

All email is not alike, and that is especially true when it comes to webmail. Different programs and systems output email in various formats, meaning the strings of metadata don't look the same. It is nearly impossible to effectively cull down a mountain of duplicative emails when the data was generated using disparate webmail and/or internal mail programs.

DSi was recently involved with a project for which we collected emails from 80 different accounts – approximately 500,000 email messages from both internal mail programs and multiple webmail applications. The emails were collected during the electronic discovery process of a case, meaning that they would need to be culled and searched by attorneys to determine what would be pertinent to the lawsuit.

Many of the emails were EML, a standard format used by multiple email programs. Generally, EML is a file extension for emails saved as MIME RFC 822, or Multi-Purpose Internet Mail

Extension with an ASCII messaging header. We also collected from other standard platforms such as Lotus Notes and Exchange, which are stored in their own unique formats. Additional emails were collected from multiple IMAP (Internet Message Access Protocol) and POP3 (Post Office Protocol) accounts, including Gmail, Hotmail (or Live Mail), Yahoo, Apple's Me.com and various others.

Each email and webmail program can structure their data differently.

The collection process showed that current electronic discovery and forensics software aren't as comprehensive as we would like to think. There is no one tool that can accurately collect all of the various email formats. Multiple methods and applications need to be employed to accommodate the myriad platforms and file types and still maintain data integrity.

There is also the issue of handling deduplication, a common way to reduce the amount of data on the front end of an eDiscovery project. The current method for the deduplication of emails is to create an MD5 or SHA1 hash of a string of text generated using portions of the emails' metadata. Because specific fields are static across different copies of the file, this is a sensible way to remove duplicate files. As an example, when you send one email to 10 different people, all the fields for "to," from," "subject," etc., are the same, meaning they are accurate to duplicate against.

However, the storage of email metadata varies greatly depending on the email system. Each email and webmail program can structure their data differently. That means different platforms may or may not contain the same fields, or, if they have the same fields, they may be named differently, or the storage of the metadata may differ, or other issues may arise.

Challenges

Challenges to be able to deduplicate different email formats:

1. Identify all the structural and metadata variations across multiple platforms.
2. Determine what needs to be modified in each program to make them all conform to each other for the purposes of hash generation.

Even though email has been around for decades, and in fact pre-dates the Internet, there is no one application that can properly deduplicate across multiple email platforms.

These issues regarding email collection and deduplication combine with existing dilemmas in the industry about file conversion methods and hashcode creation.

There are accepted processes to convert file types (ie. NSF to PST) and standardize them. They have been implemented for some time, however, the truth is that those conversion methods are flawed. Streamlining that type of process comes at a cost: Metadata can be lost – either because the metadata may be changed during conversion or because items from the original source do not get converted. That is unacceptable when those files – and the process of working with those files – can be called into a court of law.

There are accepted processes to convert file types. However, the truth is that those conversion methods are flawed.

The dilemma with email hashcodes is that, while commonly used in our industry, there is no standard method of creation. Hashcodes uniquely represent an arbitrary amount of information for the purposes of validation and verification. This data can be a file's binary information, a website password, email metadata consolidation strings or any other type of information that can be represented digitally. However, each platform has a different method of mapping data and determining the hash value. While the logic behind the algorithms is the same, the results are contrasting because the information and processes vary greatly. Differences between platforms can include the order of the fields, the delimiters between fields, or the way the data is input, such as showing time in a 12-hour versus a 24-hour clock. In the litigation technology industry, there is a need for standardizing email hashing across all processing platforms. It is not that any of their individual processes are incorrect; on the contrary, they are perfectly sound and logical. Yet, there is no way to work with that data across different platforms because there isn't a standard for how data is stored and, thus, no definitive method for hashcode creation.

Solution

So, how do you work with various email platforms?

DSi ended up designing a solution for webmail and email that performed the actions we needed to compare and deduplicate. We began by reverse engineering the various programs and then thoroughly analyzing each field from every platform to determine the differences. For example, one email platform may list attachments as "attachment1.doc; attachment2.doc",

while another might list them as “attachment1.doc,attachment2.doc”. Those slight distinctions of using a semi-colon versus a comma or not having a space will make the process of email deduplication completely ineffective.

Next, we wrote custom code to parse the data, create hashcodes and store the information so that it could be processed through a standard electronic discovery platform. The end process loads the native files, verifies metadata, changes the data temporarily to generate a hashcode, and reverts the data back to the original. The hashcode stays, which can be used to deduplicate, but the data is not changed. Additionally, we kept a forensic copy of the source files, as is customary and best practice, to compare and validate as needed.

One thing we had to consider was creating a process that would not exclude data potentially responsive to the case. As applicable, we veered on the side of inclusion versus exclusion to ensure the results were sound. We also conducted numerous quality checks, and made alterations as required, to confirm our process was accurate, effective and defensible.

ROPIN’ IN SOCIAL MEDIA

Each social media platform is different, with unique code and variations. Each one runs on its own hardware and software platform, and some, such as Facebook, have even developed custom technology to run their sites. Because of that, each requires its own method of forensically collecting data. Additionally, collection processes have to keep up with the constantly changing code base for these social media giants.

Collection processes have to keep up with the constantly changing code base for these social media giants.

Facebook

Facebook was the first platform to create a simple way to download a user’s information. The archive is comprehensive and quicker than one created with an outside solution. It includes all posts, messages and chat conversations as well as photos and videos that the user has shared. There is also the option of an “expanded archive” that includes additional historic information such as IP addresses used during logins. Facebook data is provided in an HTML format that can be viewed on a computer.

The downside of this collection module is that the user may need to download the data himself. Even if a

forensics company has the user name and password to log into the account and download the information, Facebook has implemented other security protocols that can require the account holder's participation. For example, once the email is received from Facebook noting the archive is ready for download, the link may direct you to a page with a randomly generated question that only the account holder can answer, such as naming someone in a photo. While it is possible to research the account holder and determine the answer, sometimes the most time-efficient manner is to have the individual download his own account information.

Twitter

Like Facebook, Twitter now has an easy method for users to download their own archive, which includes all of the user's tweets and retweets. The button to request your archive is under settings. Twitter will email a link to download the information. Like Facebook, Twitter's new collection module requires the account holder's participation since the link to download is sent to the email address linked to the account.

If the user is involved in downloading his own information from social media platforms, it should be in conjunction with the company handling the forensic collections.

After Twitter information has been downloaded, it needs to be displayed in a format for review by an outside party with the ability to view tweets from multiple users at one time. This can be done using the foundation of an existing application, like Tweet Nest, and modifying the code for viewing requirements. This kind of interactive web-style database allows attorneys to view and filter tweets by years, month and day, as well as search for tweets by keywords.

If the user is involved in downloading his own information from Facebook, Twitter or other social media platforms, it should be in conjunction with the company handling the forensic collections to ensure everything is handled expertly. It may need to also involve a specific protocol – i.e. that it is compressed, encrypted and uploaded to a secure FTP site.

LinkedIn

For LinkedIn, our experience suggests that the most effective way to gather data is by writing custom code. Due to the way that information is stored and structured on the site, LinkedIn is the most disjointed system of all the major social media networks and thus the most difficult

one from which to collect data. Through custom coding, we have had success in pulling all profile information, including groups to which the user belongs. At the time of writing this paper, however, LinkedIn is modifying its platform, and the upgrade may allow for easier collection.

Cloud-Based Applications

Cloud-based documents and calendars can also be collected through an existing application or by writing code to fit specific requirements. Once that information is collected, it can be converted to formats that can be opened in common programs, such as Microsoft Office.

Google's applications, such as Google Docs, Gmail, chats and other correspondence, can now be collected through their recently launched eDiscovery tool, Google Apps Vault. For a small monthly fee, Vault adds capabilities for information governance, email and chat archiving, placing legal holds, eDiscovery searching, exporting and auditing. This comprehensive suite was a needed addition for business customers, and greatly simplifies future collections of Google information.

Even if content is private, that doesn't mean that it is privileged*. Any content posted online or emailed can still be collected for a legal matter.*

WHAT DO WE HAVE THE RIGHT TO COLLECT?

While these are social media sites, there is still some expectation of privacy. The amount of privacy varies depending on the platform and how the content is distributed through it.

For example, most tweets on Twitter are public and easily accessible, but direct messages are private. Additionally, a company can't "spider out" and get information from someone just because that person is linked with the user being collected. Similarly, courts don't appreciate "friending" someone as a pretense to being able to collect that person's information.

However, even if content is private, that doesn't mean that it is *privileged*. Any content posted online or emailed can still be collected for a legal matter.

WHOA, COWBOY

Though there are many techniques for gathering data from social networks and webmail online, digital forensics and eDiscovery companies need to proceed with caution. Not every collection method is acceptable. It is important for companies to have proper authorization from the service provider. A common obstacle faced in the collection across various platforms involves the user agreement between the service provider and end user. While a forensics company can write code to collect information, doing so can violate the user agreement, and earn the negative connotation of "scraping." Each platform's terms of service should be viewed carefully to determine if the agreement will be violated – either by the manner in which collection happens or because of the information that is gathered.

It is important to have proper authorization from the service provider.

SADDLE UP

This wild west can and will be tamed. In the future, more webmail applications and social platforms will follow the leads of Facebook and Google and establish methods within the applications to collect, search and view archived records. Similarly, eDiscovery and digital forensics firms will place an emphasis upon learning and understanding the best practices involved in webmail and social media collection.

However, we are not yet to that point. Before collecting any webmail or social media, it is important to conduct an in-depth vetting process with the companies involved to learn about their procedures, protocols and quality control standards. Once these processes become standardized, we'll all ride off into the sunset.

CITATIONS

1 “It’s a Social World.” 21 Dec 2011, comScore. 13 Jul 2012. <http://www.comscore.com/Press_Events/Press_Releases/2011/12/Social_Networking_Leads_as_Top_Online_Activity_Globally>.

2 Lipsman, Andrew. “comScore Voices.” 23 Dec 2011, comScore. 13 July 2012. <http://blog.comscore.com/>.

3 “Celebrating #Twitter7.” Twitter. 21 March 2013. <http://blog.twitter.com/2013/03/celebrating-twitter7.html>.

4“Statistics.” YouTube. 29 April 2013. <http://www.youtube.com/t/press_statistics/>.

5 “How many posts are published on Wordpress.com?.” 15 May 2013, WordPress. <http://en.wordpress.com/stats/>.

6 “comScore Media Metrix.” Aug 2011, Flickr. 13 July 2012. <http://advertising.yahoo.com/article/flickr.html>.

7 “Email Statistics Report 2012-2016.” The Radicati Group, Inc. April 2012. <http://www.radicati.com/>.

8 “Key Facts.” 2013, Facebook. 23 April 2013. <http://newsroom.fb.com/content/default.aspx?NewsAreaId=22>.

9 “Celebrating #Twitter7.” Twitter. 21 March 2013. <http://blog.twitter.com/2013/03/celebrating-twitter7.html>.

10 “About.” 2013, LinkedIn. 29 April 2013. <http://press.linkedin.com/about>.

11 “About.” 29 April 2013, Tumblr. 29 April 2013. <http://www.tumblr.com/about/>.

12 “Stats.” 29 April 2013, Wordpress.com. 29 April 2013. <http://en.wordpress.com/stats/>.

13 “Start-up Pinterest wins new funding, $2.5 billion valuation.” 20 Feb 2013. Reuters. < http://www.reuters.com/article/2013/02/21/net-us-funding-pinterest-idUSBRE91K01R20130221>