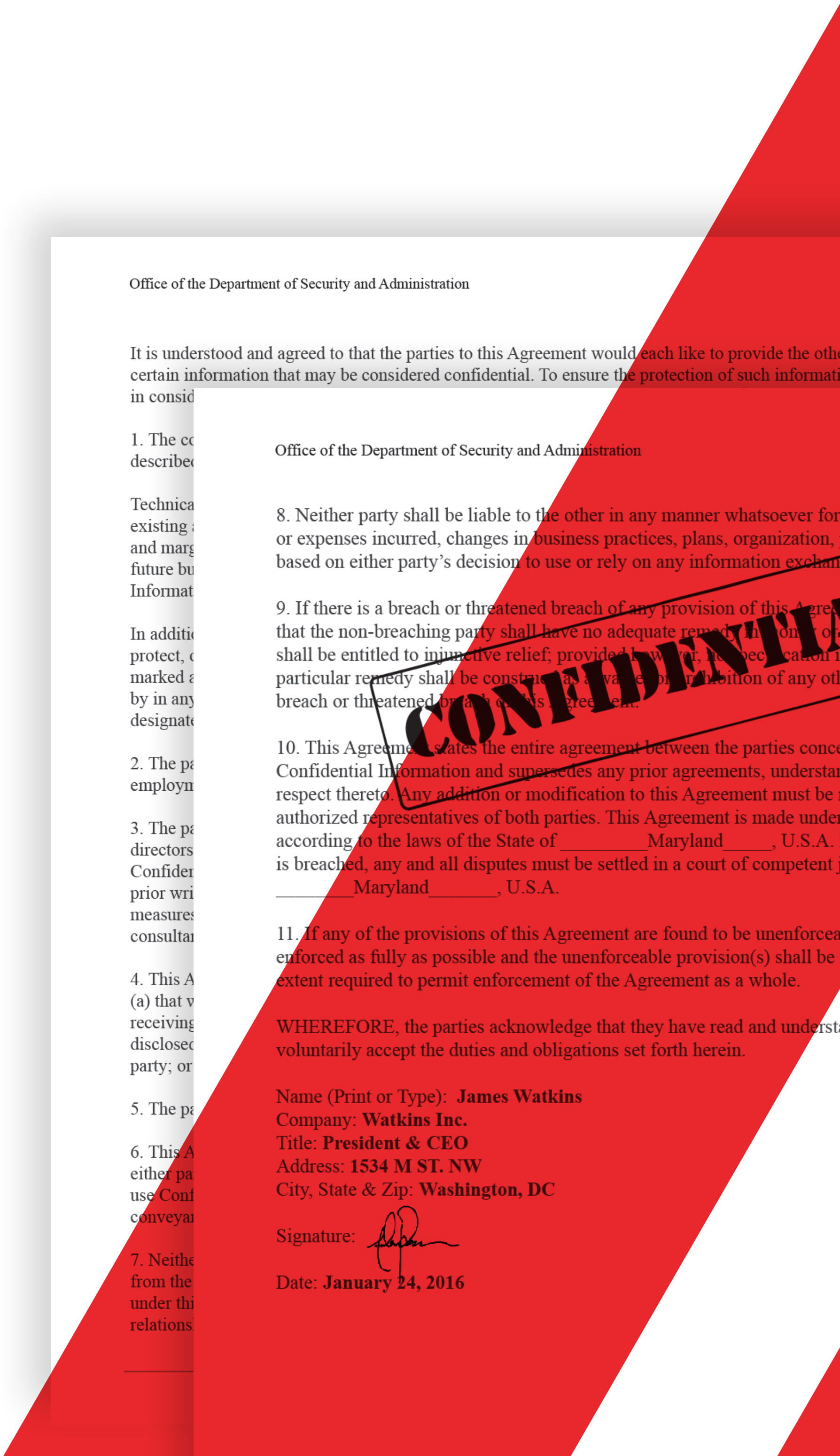
ZEROFOX®

RESEARCH

DATA LOSS PREVENTION IN A SOCIAL MEDIA WORLD

ZEROFOX RESEARCH

Philip Tully, PhD – Senior Data Scientist



Office of the Department of Security and Administration

It is understood and agreed to that the parties to this Agreement would each like to provide the othe
certain information that may be considered confidential. To ensure the protection of such informati
in consid

1. The co
described

Technica
existing
and marg
future bu
Informat

In additi
protect,
marked
by in any
designate

2. The pa
employm

3. The pa
directors
Confider
prior wri
measures
consultar

4. This A
(a) that w
receiving
disclosed
party; or

5. The pa

6. This A
either pa
use Conf
conveyar

7. Neithe
from the
under thi
relations

Office of the Department of Security and Administration

8. Neither party shall be liable to the other in any manner whatsoever for
or expenses incurred, changes in business practices, plans, organization,
based on either party's decision to use or rely on any information exchan

9. If there is a breach or threatened breach of any provision of this Agre
that the non-breaching party shall have no adequate remedy in money or
shall be entitled to injunctive relief; provided however, no specification i
particular remedy shall be construed as a waiver or prohibition of any oth
breach or threatened breach of this agreement.

CONFIDENTIA

10. This Agreement states the entire agreement between the parties conce
Confidential Information and supersedes any prior agreements, understan
respect thereto. Any addition or modification to this Agreement must be
authorized representatives of both parties. This Agreement is made under
according to the laws of the State of ________Maryland____, U.S.A.
is breached, any and all disputes must be settled in a court of competent
________Maryland_______, U.S.A.

11. If any of the provisions of this Agreement are found to be unenforcea
enforced as fully as possible and the unenforceable provision(s) shall be
extent required to permit enforcement of the Agreement as a whole.

WHEREFORE, the parties acknowledge that they have read and understa
voluntarily accept the duties and obligations set forth herein.

Name (Print or Type): **James Watkins**
Company: **Watkins Inc.**
Title: **President & CEO**
Address: **1534 M ST. NW**
City, State & Zip: **Washington, DC**

Signature:

Date: **January 24, 2016**

Social media has become the new battleground for data exfiltration due to its wide attack surface, lack of controls and inherent user trust. ZeroFOX Research provides in depth analysis of the adversarial tactics, including a 7 year timeline of major events, and delivers recommendations to modernize security best practices for social media and Data Loss Prevention (DLP).

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The rise of social media has ushered in an era of rapid and widespread access to information, but these benefits come with new liabilities. Digital communication is as seamless as it's ever been for malicious actors, opening up the door to a host of new methods of exfiltrating data outside the confines of the modern organization's security perimeter. The diverse and continuously evolving nature of social information severely complicates forming an effective response to these emerging threats.

Social media is rife with issues, from an employee inadvertently revealing sensitive information to the insider threat absconding with competitive intel to an external data breach involving a nation state actor. Social media has changed the game when it comes to data loss, both in terms of exposure and sheer vastness, making violations far more difficult to detect. This is further complicated by emerging and ever-changing social attack vectors, such as malware and phishing, which occur outside of corporate network defenses. In this report, ZeroFOX Research details the many social media data loss risks and threats and outlines a multi-layered approach that security teams can adopt to detect and prevent this data loss.

HIGHLIGHTS:

- Detailed synopsis of how data exfiltration can be carried out through social media
- 2 year analysis of hundreds of millions of social media posts demonstrating proliferation of PII disclosure
- Timeline of major data exfiltrating events performed using social media over the past 7 years
- Examples of social media data loss and how they can be prevented, detected and remediated
- Recommendations to modernize security best practices for individuals and organizations

TIMELINE OF MALICIOUS SOCIAL MEDIA DATA EXFILTRATION EVENTS

SVELTA
August 2009

@upd4t3 and @Botn3tControl: C&C malware searched Twitter and Tumblr for Base64-encoded URLs or executables to download, run and steal banking credentials in Brazil [1].

TROJAN.WHITEWELL
September 2009

Malware used Facebook status updates to coordinate C&C servers. Malicious scripts parsed HTML code, extracted and inserted data; used the account to fetch contact URLs and execute commands through remote URLs [2].

UNMASK PARASITES
November 2009

Used Twitter API to pass malicious Javascript functions as callback parameters, accepting trends as input. Generated new domain names based on the second character in the most popular trend two days prior that obfuscated malicious iframe injection [3].

2. PLUGGING THE LEAK IN SOCIAL MEDIA FOR ENTERPRISE SECURITY

Data exfiltration (i.e. data loss, data extrusion, data leakage) is the unauthorized transmission of sensitive information from inside a privileged access point. Because it can closely resemble the normal flow of data traffic, it is very difficult to detect and 'right' the sinking ship. Traditionally viewed in the context of the network, endpoint or email, data exfiltration is a known issue that can result in huge financial and reputational losses for victimized organizations and individuals. But when it comes to social media, security practitioners are increasingly finding themselves awash in a deluge of OSINT data.

Social media is a formidable attack surface due to its sheer size and breadth (Table 1). With ever-increasing volumes of data being poured into these different networks, detecting data-exfiltrating posts can be like finding a needle at the bottom of the ocean. The tides have shifted even for the largest and most talented security teams, as it's become impossible for humans to navigate through this information to identify potentially harmful threats. What's more, social media poses additional unique risks not typically encountered on traditional points of access like email – there's a torrent of different mechanisms such as #hashtags, @mentions and lists for users to instantly broadcast data to expansive global audiences.

Network	Posts per day	Monthly users
Facebook	1b	1.55b
Twitter	500m	320m
Instagram	95m	400m
YouTube	300k	1b
Google+	530k	540m
LinkedIn	18.5k	100m
Pinterest	2m	100m
Pastebin	20k	17m

TABLE 1.
Social media's broad attack surface and audience exposure size.

THE SHADOW CLOUD

April 2010

A complex cloud network leveraged Twitter and other social networks as disposable C&C locations to control target systems. China exploited reputations of these 3rd party sites to avoid detection and execute elaborate cyber espionage campaigns [4].

TWITTERNET BUILDER

May 2010

Researchers provided a tool for creating malware that used Twitter as its C&C channel for posting and distributing URLs that deliver malicious payloads to the victim [5].

BYZANTINE HADES

June 2010

Malware used Facebook as its C&C channel by posting to the account page that would subsequently respond with implant commands [6].

Social media also lacks the kind of industry security precedent that a platform like email has developed after weathering wave after wave of high-profile attacks. This should be troubling given that social media is much more trusted than email: only 11% of users open unknown emails and 22% open attachments in unsolicited emails, yet 36% of users accept unknown friend requests and 27% check their social media accounts at work before their emails.[1] This inherent trust is typically an issue of perception; users commonly take the trust associated with their friends, families and other connections and mistakenly anchor it alongside their trust in social network platforms themselves.

Social media is even more exposed than email since it's outside of an organization's standard network security protocols. It gives the attackers a cost-effective, user-friendly way to organize their data loss campaigns and ensure their malicious objectives can be achieved. It comes as no surprise that organizations both large and small are woefully under-equipped to address DLP when it comes to social media. The security industry readily admits these shortcomings as well, with a shocking 79% of industry survey respondents describing that security processes for Internet and social media monitoring simply don't exist, or are partially or inconsistently deployed.[2] An additional 43% of fraud prevention managers and IT directors recently reported that employee access to social media websites and services is their biggest obstacle when it comes to DLP.[3]

Without blocking access to the social network websites themselves, which is an increasingly unpopular option for the socially engaged and connected workforce, the organization is faced with tackling these murky waters head-on. However, organizations aren't defenseless, and there are concrete steps that can be taken to protect and remediate data loss through social media. ZeroFOX has pioneered machine learning technology that automatically alerts our customers to inbound threats that could cause data loss through social channels. Alerts are also generated on outbound data that's already been exfiltrated in order to minimize damages and costs of the post after-the-fact.

In the margins of this whitepaper we provide a timeline of the major events over the past 7 years in which social media was used as a covert channel to transmit data from within an internal network.

ZeroFOX's technology is a multi-layered monitoring solution that provides a beachhead for businesses to shore up their visibility of and control over DLP on social media. **Fig 1** outlines three different ways that data loss can occur through social media, which will inform the proceeding structure of this white paper.

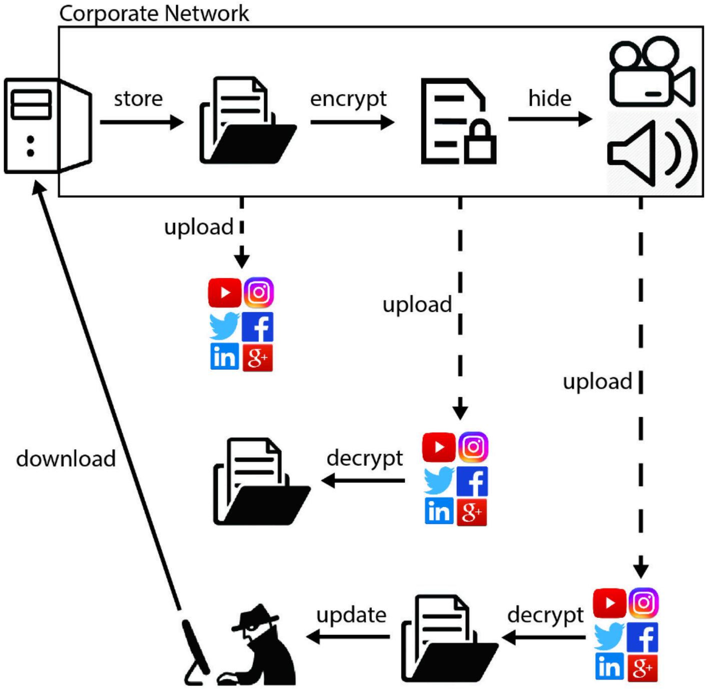


FIGURE 1.
Social media introduces new exit points for data exfiltration. Three examples illustrated above are denoted by dashed vertical arrows. From left to right, at a high level, we identify 1) Inadvertent data loss involving sensitive information posted directly to the social network, and more intentional forms of data loss like 2) The insider threat involving a disgruntled employee divulging company secrets through encoded social channel data, and 3) External data breaches by bad actors looking to hack into the corporate network and establish Command and Control (C&C) to maintain their data siphon.

LINKEDOUT

April 2012

LinkedIn iOS App was found collecting full meeting details from mobile iOS calendars, and sending out the subject, location, meeting time and personal meeting notes from personal and corporate calendar accounts in plaintext to their own servers [7].

BACKDOOR.MAKADOCS

November 2012

Google Docs, YouTube and Facebook were identified as easy targets for data exfiltration where SSL obfuscates what content is being exfiltrated. Use of encryption tools like TrueCrypt further evade detection [8].

MINIDUKE

February 2013

Allegedly Russian, malicious PDF spread through email with malware callbacks using a URL found via Twitter search. Tweets encoded to disguise URLs to download malware; then connected C&C server delivers payloads disguised as GIF images [9].

3. INADVERTENT DATA LOSS

3.1 PII EXPOSURE ON SOCIAL MEDIA

According to the NIST SP 800-122 guidelines,[4] personally identifiable information (PII) is defined as:

> *"...any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial and employment information."*

These specific examples are explicitly listed as PII:

- Name, such as full name, maiden name, mother's maiden name, or alias.
- Personal identification number, such as social security number (SSN), passport number, driver's license number, taxpayer identification number, or financial account or credit card number.
- Address information, such as street address or email address.
- Personal characteristics, including photographic image (especially of face or other identifying characteristic), fingerprints, handwriting, or other biometric data (e.g., retina scan, voice signature, facial geometry).
- Information about an individual that is linked or linkable to one of the above (e.g., date of birth, place of birth, race, religion, weight, activities, geographical indicators, employment information, medical information, education information, financial information).

Unauthorized access, use or disclosure of PII can have damaging effects on individuals and organizations alike. PII can be used against individuals for the purposes of identity theft, blackmail or embarrassment. For the organization, it can harm public relations and result in legal costs.[5]

Unlike traditional enterprise software, social networks systematically encourage the public disclosure of PII as part of the "community." People are hard-wired to maximize likes, shares, +1s, endorsements, retweets, repins and upvotes. In doing so, they inadvertently overexpose potentially sensitive data on a massive scale. In fact, business models of the networks themselves are explicitly predicated upon behavior like this; revenue generated from their user-tailored advertising annually exceeds tens of billions of dollars and growing. At the same time, security implications of readily accessible personal data are often overlooked and underestimated.

JANICAB
July 2013

Malware continuously captures screenshots and recorded audio for YouTube upload, then used C&C server info embedded in video's comments to execute new commands [10].

SKYHIGH
March 2014

Single IP address with data-exfiltrating malware on Twitter generated over 100,000 tweets per day leaking sensitive financial institution info. A single energy company malware-infected computer made 3.8 million attempts to exfiltrate data via Facebook [11].

TRIPWIRE
November 2014

A Fortune 500 company had sensitive data exfiltrated from their network through YouTube video uploads with embedded steganography [12].

ONIONDUKE

November 2014

Malware searched within configuration data for Twitter account names and tweeted with links to image files embedded with malware [13].

F0XY

January 2015

The trojan downloader f0xy used VKontakte to obtain the address of its primary dynamic C&C server before the infected box was utilized to mine cryptocurrency [14].

HAMMERTOSS

July 2015

Exploited by APT29 (Russia), malware scanned posts for hashtag and image links in new Twitter handles generated daily. For each image, Malware used a steganographic decoder on image to extract and execute malicious C&C instructions [15].

Data is the fundamental currency of the Information Age, and if information is power, social media is an OSINT goldmine for would-be bad actors. Absent any explicit social post containing sensitive data, a person's social media account footprint already gives the adversary a leg up. Though the field is optional, social media users frequently personalize their account pages by sprinkling in specific details into their descriptions or "About Me" fields including job titles, place of employment, hobbies, interests, and hyperlinks to external websites or personal blogs. Other fields may contain indicators like location via check-ins and other geographical indicators, date of birth, composition of friends and followers, first and last names, email address and demographic identifiers like gender, weight, race and religion. Even personal telephone numbers are often divulged, ironically the most common element of PII that is used for 2-factor authentication.

Other commonly leaked information that is heavily sought after includes login credentials, IP addresses, cryptographic keys or passwords, proprietary information, intellectual property and secrets, design diagrams, source code, AWS keys on GitHub, tokens, subdomains, domain history and legacy portals, router IDs, x-ray photos, and other protected health information (PHI). A comprehensive overview of PII data gathered from a subset of social media users over the past 2 years illustrates the proliferation of this type of exposure (**Fig 2**). One publicly available tool even advertises its ability to flush out such target-related information from public data sources.[6]

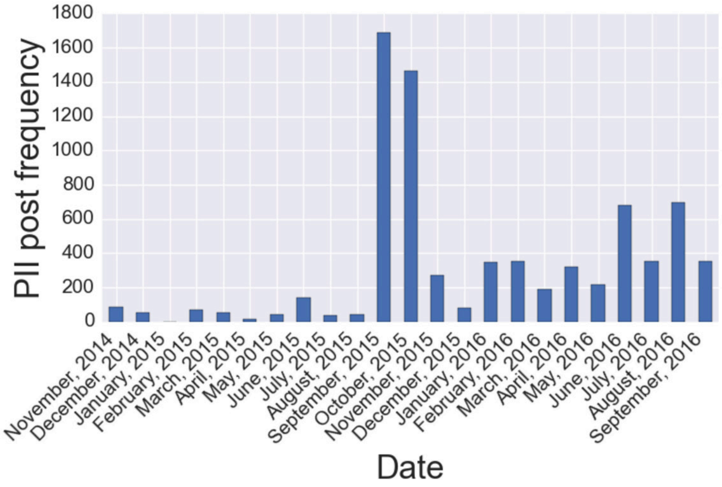


FIGURE 2.

A 2-year analysis of 624,021,944 posts demonstrates the prevalence of PII disclosure on social media.

SNEAKY CREEPER
July 2015

Data exfiltration framework that used Twitter,

Tumblr and SoundCloud. The messages used RSA and Base64 encoding while incorporating steganography to hide data within audio and images [16]

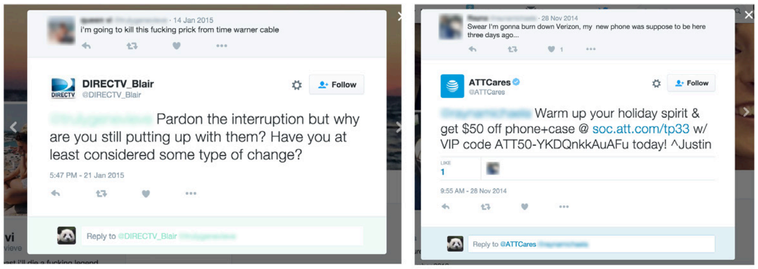
C&C-AS-A-SERVICE
September 2015

The Dukes exploited Twitter to communicate with infected machines, using Microsoft OneDrive as C&C infrastructure to exploit corporate network access. Attackers then used Twitter to communicate with infected machines [17].

TWITTOR:
November 2015

A botnet C&C infrastructure used the Twitter API to push direct messages, which unlike tweets allow character length-unbound messages [18].

People's particular job functions can be successfully identified in just a few clicks, and organizational sketches are easily crafted to identify the most susceptible and high-value targets. Employees and customers regularly reveal their relationship to organizations, whether it be on LinkedIn or through a complaining social post about a bad service or product. Disdain for marketing campaigns and dissatisfaction with brands and products can lead to social poaching (**Fig 3**), or a phenomenon in which corporate competitors subsequently swoop in and try to scavenge upset customers. Customer self-identification can lead to other problems for organizations including targeted scams and fraudulent activity.



[redacted] · 14 Jan 2015
i'm going to kill this fucking prick from time warner cable

DIRECTV_Blair
@DIRECTV_Blair

[redacted] Pardon the interruption but why are you still putting up with them? Have you at least considered some type of change?

5:47 PM - 21 Jan 2015

Reply to @DIRECTV_Blair [redacted]

[redacted] · 28 Nov 2014
Swear I'm gonna burn down Verizon, my new phone was suppose to be here three days ago...

ATTCares
@ATTCares

[redacted] Warm up your holiday spirit & get $50 off phone+case @ soc.att.com/tp33 w/ VIP code ATT50-YKDQnkkAuAFu today! ^Justin

LIKE
1

9:55 AM - 28 Nov 2014

Reply to @ATTCares [redacted]

FIGURE 3.
Two examples of social poaching, where upset customers who divulge their service provider are targeted by competitors that seek their business and helpfully provide a service alternative.

3.2 ACCIDENTAL DATA EXFILTRATION

While the previous examples emphasized the problem of generic PII being corralled by an adversary who could start to infer missing pieces of a puzzle, a more direct and careless form of data exfiltration also exists. Social media users regularly post extremely sensitive information and broadcast it for the world to see on social media.

Consider users who post pictures of their debit or credit cards onto social media, allowing others to make purchases on their card immediately. Believe it or not, this behavior is rampant. The the Twitter user @NeedADebitCard (**Fig 4**) actively retweets pictures of their users debit and credit cards, exposing it to their 18,000 followers at the click of a button. While seemingly taking a playful approach to educating the public about this risky behavior, this user has likely caused massive headaches for nearly 200 victims and their credit card issuers.

BLACKENERGY2
November 2015

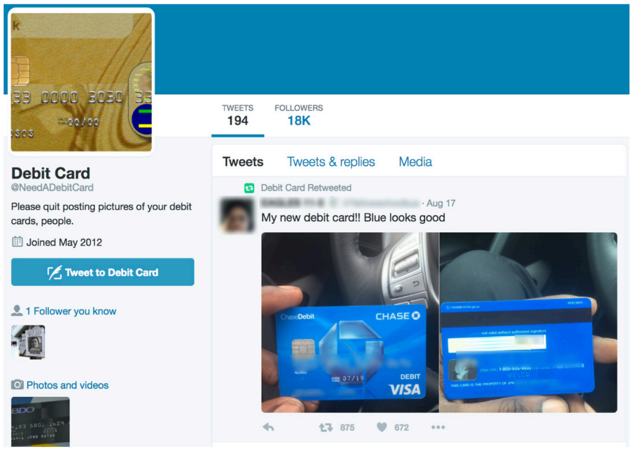
Plugin used PNG files to interact with Google+ over HTTPS. Leveraged the OLE stream Windows structured storage API and GDI+ bitmap functions to ingest a target PNG file and regenerate with RC4-encrypted data containing malicious C&C config [19].

DATA EXFILTRATION TOOLKIT
March 2016

Publicly released tool implements data exfiltration through Flickr, Twitter and YouTube. One or more channels could be used in each iteration [20].

INSTEGOGRAM
August 2016

A proof of concept hid messages within Instagram images. C&C trojan remotely executed malware using a steganographic decoder to extract and execute payloads from each image [21].



Debit Card
@NeedADebitCard

Please quit posting pictures of your debit cards, people.

Joined May 2012

Debit Card Retweeted

[redacted] · Aug 17

My new debit card!! Blue looks good

FIGURE 4.
@NeedADebitCard trolls fellow users who tweet pictures of their debit and credit cards.

Yet another accidental form of data exfiltration involves publicly posted pictures. In the summer of 2015, British Health Secretary Jeremy Hunt accidentally breached patient confidentiality by posting a photo of himself posing with physicians in front of a board listing patient information to his 70,000 followers.[7] This and similar disclosures[8] are all-too-common occurrences for employees who like to take selfies at the workplace, which can often display sensitive organizational information like product roadmaps, architecture diagrams, software stacks, or customer information. In addition to data loss, culprits can unknowingly violate industry-wide compliance mandates, potentially resulting in hefty financial penalties for the organization in question. Accidental data exfiltration can even pose national security risks: active service members have previously posted patrol times, details of sensitive visits and photos of restricted areas.[9] In 2015, the goeloactions tags on a Russian soldier's social media posts revealed that the Russian army had crossed the border into Ukraine.[10]

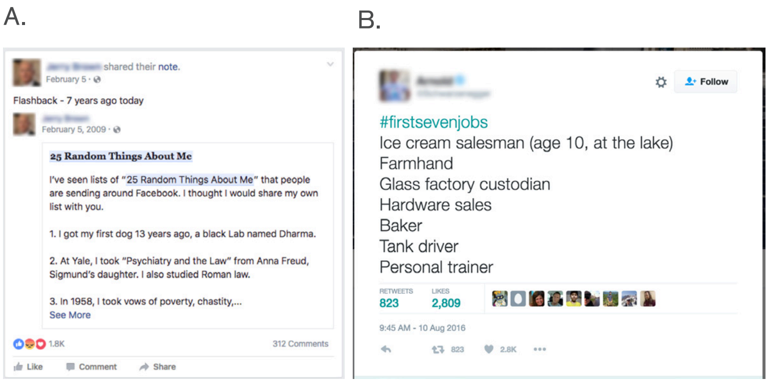
There's a strong incentive to post pictures to social media while on vacation or at a popular event; you've traveled to an exotic place after months of hard work and want to share with friends and family the hard-earned fruits of your labor. But there are often unforeseen consequences: event ticket barcodes can be copied and resold for a higher value or used directly, disallowing access to the true purchaser.[11] Disclosing that you're not home implicitly announces your absence, essentially inviting burglars over for an easy score. Even worse, if you're especially flashy in your historical social media presence, as in posting luxury goods or lavish destinations, you make for a desirable target. Insurance companies are beginning to catch on, invalidating claims of customers they don't perceive as having been reasonably careful in protecting their possessions.[12]

CRYLOCKER
September 2016

Social Engineering ransomware that locked key local files. CryLocker exfiltrated PC config data hidden in PNG image uploaded Imgur. A unique filename would be created to call back to the C&C server to alert to new takeover [22].

A popular meme spread across Facebook in which users voluntarily disclosed "25 random things" about themselves, which included the type of information that password reset questions are based on (Fig 5A). More recently, users divulged their first seven jobs, which quickly blossomed into a viral social media trend (Fig 5B). The trend invited onlookers to gather public PII, all of which was available through simple hashtag search queries. Many companies, like banks, credit card issuers and insurance agents, ask their customers to answer security questions having to do with previous employment as a protective mechanism. The practice is also commonly employed in such everyday identity-sensitive tasks like driver's license registrations, automobile purchases, mobile phone plan sign-ups, and student loan applications.

Finally, there are those users who simply make a high stakes blunder by posting something intended only for private eyes. This has happened to one of Instagram's most followed users, Scott Disick,[13] and the Twitter CFO, Anthony Noto.[14] Noto posted extremely sensitive comments about buying another company, information meant for a direct message.



A.

[redacted] shared their note.
February 5 ·

Flashback - 7 years ago today

[redacted]
February 5, 2009 ·

25 Random Things About Me

I've seen lists of "25 Random Things About Me" that people are sending around Facebook. I thought I would share my own list with you.

1. I got my first dog 13 years ago, a black Lab named Dharma.

2. At Yale, I took "Psychiatry and the Law" from Anna Freud, Sigmund's daughter. I also studied Roman law.

3. In 1958, I took vows of poverty, chastity,...
See More

B.

[redacted]

#firstsevenjobs
Ice cream salesman (age 10, at the lake)
Farmhand
Glass factory custodian
Hardware sales
Baker
Tank driver
Personal trainer

RETWEETS 823 LIKES 2,809

9:45 AM - 10 Aug 2016

FIGURE 5.
The A) "25 Random Things" Facebook and B) #firstsevenjobs from Twitter trends were ripe sources of PII.

4. INTENTIONAL DATA EXFILTRATION

4.1 EXTERNAL DATA BREACHES

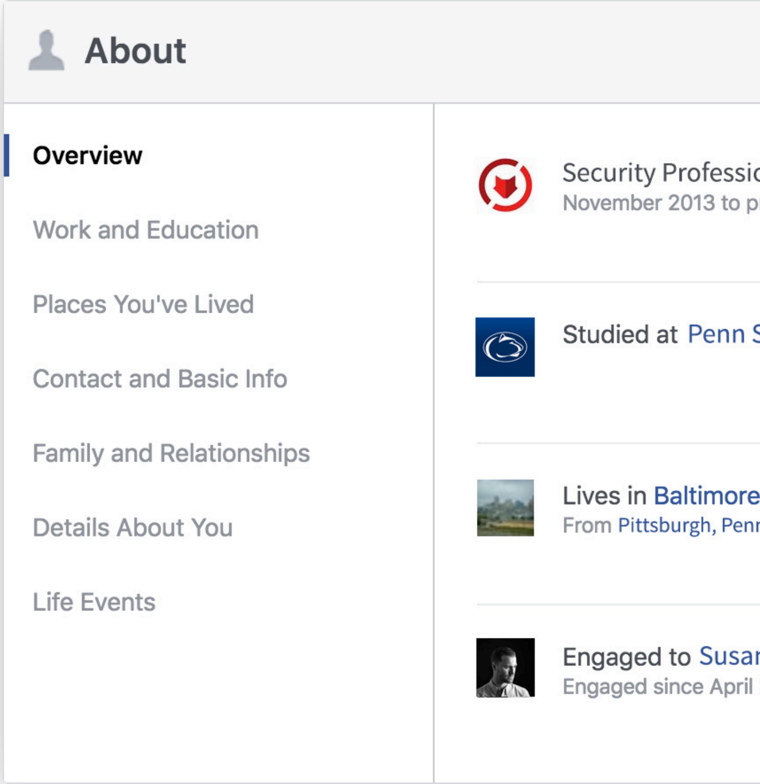
Levaraging PII for Social Engineering

It is easy enough to navigate to a target's account to record PII, and nefarious adversaries can perform more sophisticated reconnaissance by extracting information across several different social networks. They each offer some level of search functionality that permits manual retrieval and identification of accounts. Furthermore, many PII-laden fields are programmatically accessible through their public APIs, opening up the frightening possibility of automating the harvesting of a target's data. Even when the data isn't API-accessible, ambitious adversaries can circumvent these processes to anonymously vacuum data right off the accounts themselves through programmatic scraping, permitted loose privacy settings are configured by the target account.

Rather than indirectly extracting PII from the social networks, adversaries can alternatively use that PII for the purposes of social engineering by actively publishing extracted data back across the networks. Social media is no stranger to this type of behavior, as it's been shown that 39% of all social engineering attacks take place in this manner.[15]

But how might this scenario pan out? First, this is the spear phisher's dream; the insight gained from publicly available social media PII can be used to handcraft tailored messages geared towards the target's interests, and contain messaging that would make them likely to click any attached link, irrespective of any previous interaction. Facebook direct messages have been previously used for this purpose.[16] On Twitter, users click on links within replies from users they are neither followed by nor follow up to 66% of the time, given they are targeted with clickbait tweets containing near and dear topics.[17]

Once a click-through is achieved, the spear phisher can use tools like MetaSploit[18] to generate fake login websites, mimicking the social network itself for example, while actually rerouting to third party websites capable of stealing input credentials. After being hijacked, these accounts can be utilized to extract more PII through engagement with the target's connections. One can then more easily hijack additional accounts, rinse and repeat. This type of fake familiarity is all too common, too. In fact, 20% of all phishing attacks are perpetrated through social media[19] and 54.4% of social media users report receiving phishing attacks.[20] The total cost to organizations of phishing adds up to approximately $1.2 billion annually.[21]



Security Professio
November 2013 to p

Studied at Penn S

Lives in Baltimore
From Pittsburgh, Penn

Engaged to Susan
Engaged since April

A second tool in the social engineer's war chest is the creation of an impersonation account. An impersonation account is common first step in a social engineering attack, both in the information gathering phase and the attack phase. Ironically, a recent example showed that security industry members made for especially good targets.[22] These impersonating accounts often including copy-and-pasted fields from the victim's account and can incorporate other credibility building tactics like purchased followers. This is trivial to carry out, as social network followers can be bought in bulk through third party websites.[23] The networks have tried to blunt this behavior by introducing verified account tags to sort out the good guys from the bad, but it's not a silver bullet; tech titans recently teamed up and found scammers impersonating 2,400 different legitimate businesses.[24] ZeroFOX's own analyses lends support to these trends, with a recent sample of approximately 100 organizations showing more than 1,000 impersonation accounts being created on a weekly basis.[25]

The ramifications of these tactics can be extremely harmful. The social engineer's coup de grâce is to hijack the target's actually identity altogether. Leveraging the aforementioned social engineering techniques, a perpetrator might go as far as carrying identity theft out in real life once enough information has been absconded. Documented cases have involved everything from accruing vast monetary debt to committing crimes in the victim's name. Losses don't necessarily need to be financial, either. Victims are often tasked with arduous, long-term problems like trying to restore their personal reputations or correct misinformation. The US Federal Trade Commission tallied 9.2 million online victims who had personal data stolen by cyber criminals last year, with over 19 people falling victim to identity theft each and every minute.

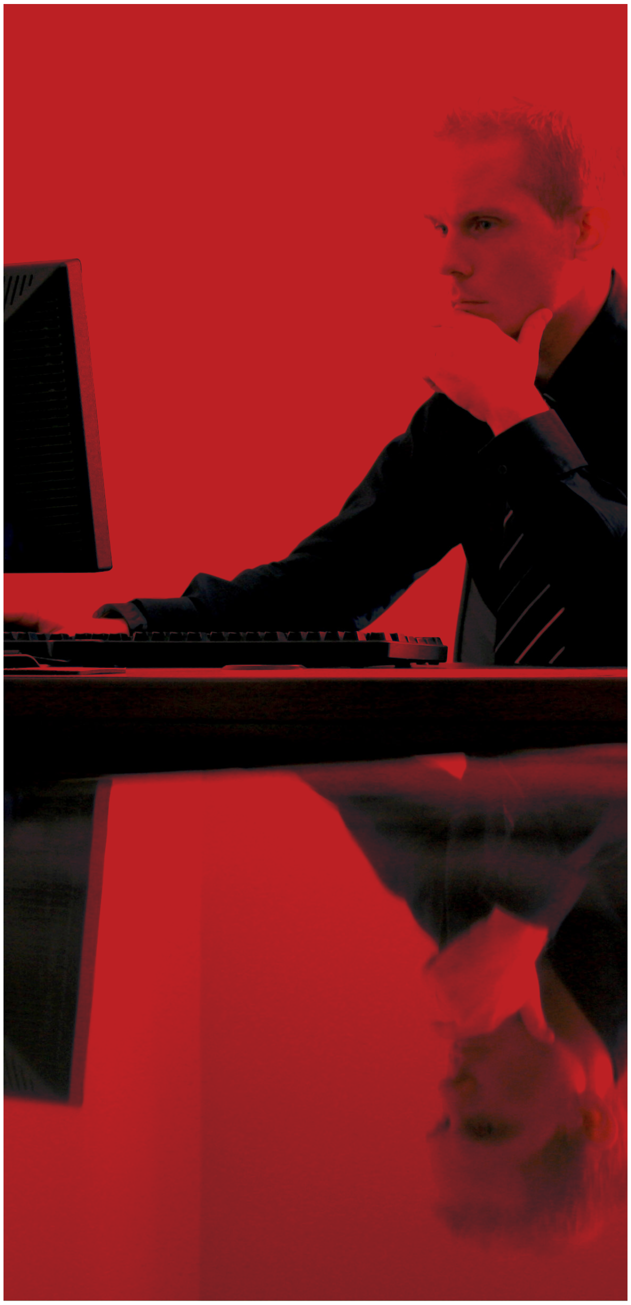
Extortion is another popular route to profit off victim data. Typically, criminals will contact their victims with a ransom that's necessary to withhold the sensitive information. Failing to comply can result in doxxing, the process of releasing PII or other sensitive data through a public arena like social media. As such, not only can social media be an effective tool for harvesting dox-worthy data, but it can simultaneously be used as as a distribution mechanism to complete the attack. It's a one-stop-doxxing-shop, as it were.

Launching an automated script to exfiltrate data

More technically adept adversaries have other tools at their disposal, including creating malware delivery algorithms that compromise host networks. Reasons for doing so include, but are not limited to, nation state or corporate espionage, organized crime, and hacktivism. As it turns out, 69% of breached organizations found out they suffered external breaches after the fact.[26] Surveys of enterprises and financial institutions indicate that they incur external attacks more than once per month, and spent an average of $3.5 million to remediate residual damaging effects.[27]

When it comes to social media, attacks do not typically exploit vulnerabilities or 0-days within the platforms, rather they try to exploit platforms for purposes not originally intended to serve. Usually, attackers try to take advantage of the normal behavior of target systems in order to remain disguised and maintain control for extended periods of time. For example, C&C involves URLs that can be continually and remotely updated to point to C&C servers. When connection to the C&C servers is blocked at the firewall level, compromised computers within the network can remain controlled because websites like social networks and microblogs are trusted and commonly overlooked by network administrators. Remote URL updating through these unblocked intermediaries ensures new C&C servers can maintain contact with the infected internal computer.

4.2 THE INDSIDER THREAT



Both the strongest asset and, simultaneously, weakest link of any organization is its own employees. Inevitably, workers gain access to privileged and sensitive information systems, including personal, customer, and employee data. Executives and managers trust their workers to perform honorably and carry out critical tasks in the best interest of the business. But this isn't always the case; disgruntled employees who are on their way out but still have access to old system credentials can create back doors before having their accounts be deprovisioned, or sell information to competitors or nefarious groups. Insider threat attacks cost organizations an average of $445,000 per instance to remediate after the fact.[28]

The simplest version of this attack can occur when an innocuous office member, such as a temporary worker, shoulder-surfs (peeks over the shoulder of a screen-glued employee), or even approaches a vacant desk to view a computer screen that hasn't been locked. In today's world of mobile devices, it's easy enough to photograph sensitive documentation, computer screens, or whiteboards. This type of "visual hacking" was found to be successful in logging sensitive information 91% of the time.[29]

The social networks themselves have huge targets on their backs, too. After all, these organizations are treasure troves with unfettered access to the type of high value data we've outlined thus far. In fact, there is alarming precedent for this too. Major data leaks have successfully been executed against the likes of LinkedIn, Twitter, VK, Tumblr, MySpace, and most recently DropBox.[30] These attacks typically involve a significant number of user passwords being exfiltrated from the social network databases, and subsequently doxxed or sold for hefty prices on the dark web, or used to steal accounts from other online assets the victims have unadvisedly protected with identical passwords.[31]

5. PREVENTION, DETECTION AND REMEDIATION

5.1 PREVENTION

The first line of defense for DLP involves detecting the threat. Incoming social media posts that contain phishing or malware links can be analyzed, alerted on, classified as malicious, and blocked or taken down within moments of hitting the social networks. All of this can occur faster than the time it takes for key employees to check their notifications and click on the URL. This can then be consumed into perimeter, endpoint protections, and SIEMs. Such control can help protect against intentional data exfiltration (see Section 4).

To compliment this, an organization should consider instituting a social penetration testing program to ascertain the vulnerability posed by their own employees and/or customers. The insider threat can be prevented by footprinting applicants' social media profiles during the employee hiring process. There may be publicly-facing posts of interactions that could raise red flags and indicate concerning behavioral patterns, enabling the organization to prevent expensive hiring mistakes in the long run.

5.2 DETECTION

But what happens when the attack has been successfully carried out and sensitive data is continually being exfiltrated over social media? Just as incoming posts containing malicious content can be alerted on, so can outgoing posts containing sensitive data.
This turns out to be quite difficult in practice for several different reasons, including, in no small part, the sheer scale of the social media dataset (see Table 1). The social footprint of an organization goes beyond the company account or official pages, or even what's typically available from off-the-shelf network search APIs. The attack surface is porous too. It can include the use of hashtags with organizational branding and the organization's employees or customers accounts among many other exit points. Furthermore, social media syntax is laden with syntactical oddities like abbreviations, #hashtags, @mentions, and slang, making it difficult to apply previously established techniques that excel on traditional text corpora.

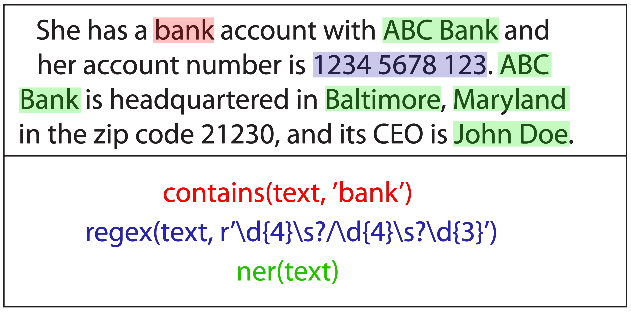
Several techniques to consider for matching patterns within social media posts are available through the ZeroFOX platform, including:

1. Exact string matching to isolate posts containing specific words; rules can be written to generate alerts based on the presence of the word "bank" within incoming social media posts (Fig 6). The unfiltered presence of specific words alongside other words signifies compliance infractions like FFIEC, FINRA, HIPAA, and PCI standards to name a few, leading to ensuing penalties and high legal fees.

2. Regular Expressions (i.e. Regex) are sequences of characters that encode patterns to capture within an input string (Fig 6). Regex is powerful since it utilizes a fraction of the characters compared to a full enumeration of all possible variations of the desired result. It can help capture more complex, higher level patterns like street addresses. Addresses have many syntactical variations, like "St." instead of "Street", "Twelve" instead of "12" or even extra intervening whitespace or punctuation that would not match the query string verbatim. Exfiltrated data is often unstructured and highly variable, and exact matches on a simple strings are prone to error accumulation. Regex can be used to programmatically uncover more nuanced plaintext in formats like base64, XOR or hexadecimal, which can encode sensitive data and C&C commands (see Section 4).

Continued on next page...

These techniques apply mostly to detecting inadvertent data loss and intentional external data breaches. But in terms of the insider threat, existing employees' social media activity might reveal information that can be used in follow-up legal proceedings in cases where fraud has been perpetrated. Fraud might be something overtly malicious, such as the intention to exfiltrate data, but can alternatively be as straightforward as identifying staff that abuse long-term sick leave under suspicious circumstances or give explanations surrounding workplace accidents.[32]



She has a bank account with ABC Bank and her account number is 1234 5678 123. ABC Bank is headquartered in Baltimore, Maryland in the zip code 21230, and its CEO is John Doe.

```
contains(text, 'bank')
regex(text, r'\d{4}\s?/\d{4}\s?\d{3}')
ner(text)
```

FIGURE 6.
Applying increasingly complex natural language processing methods to post content can help reduce false positives and false negatives when detecting instances of data exfiltration. The pseudocode in the bottom panel corresponds to techniques1-3 from the sidebar, and the top panel indicates color-corresponding matches resulting from the rules applied in the bottom panel.

3. Named Entity Recognition (i.e. NER, Fig 6) identifies named entities within an input string. Named entities can be names of people, organizations, locations, dates, or quantities like time or money. Trained by supervised machine learning, NER takes into account each word and neighboring ones to optionally label them as named entities. The models incorporate statistical evidence from an annotated training corpus of labeled sentences, allowing NER to generalize previously unseen sentences and word combinations by weighing natural language features like capitalization, sentence context, punctuation, and part of speech.

4. Bot classification can identify accounts that automatically post. Bots could be performing scheduled updates and maintenance to connect with C&C servers, a telltale sign of some of the most prominent data-exfiltrating malware (**see the Timeline**). Social media bot accounts exhibit common behavioral features like the presence of URLs within posts, unconfigured default fields like background image and description, a high ratio of following to followers, consistent source of the application posts are made from, short account ages, and little variation in the timing between posts or the content of each post compared to a typical user.

5. Optical Character Recognition (OCR) can be applied when text-based approaches are not enough. When text is explicitly superimposed on an image, OCR can extract that text, which can subsequently be analyzed by some of the techniques described in 1-3 above.

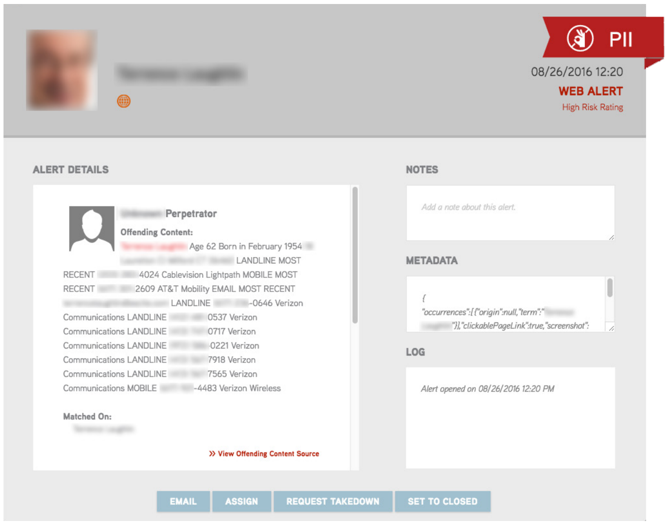
6. Steganalysis refers to the statistical determination of irregularities within images, including alterations that may contain sensitive data or a malicious payload. In most cases, the hidden messages are encrypted.

5.3 REMEDIATION

Traditional security measures emphasize detection of elevated email or other incoming traffic occurring within the organization's network perimeter. But social media operates outside of enterprise control, with content that only the social networks themselves have authority to take down. Social networks typically respond quickly to requests to take down content that violates their Terms of Service. What's lacking is a monitoring and visibility mechanism to help organizations parse through all the harmless noise and identify meaningful and actionable risky content to pass along for remediation.

The ZeroFOX Platform addresses these concerns and more. By ingesting data from LinkedIn, Facebook, Twitter, Instagram, Youtube, Google+, Tumblr, Reddit and more, it continuously applies prepackaged and tuned rules to social media accounts and posts in order to identify meaningful and actionable content.

Becuase incoming social media data is streamed through the ZeroFOX Platform and the different techniques noted above can be applied to alert on PII (**Fig 7**), relevant risky social media data can be identified in a matter of seconds. Compare this to recently published statistics on how long it takes organizations to detect data exfiltration: less than 2% within 24 hours and less than 3% in the first week. The average time for detection is on the order of months.[33]



[redacted]

08/26/2016 12:20
WEB ALERT
High Risk Rating

[redacted] Perpetrator
Offending Content:
[redacted] Age 62 Born in February 1954 [redacted]
[redacted] LANDLINE MOST
RECENT [redacted] 4024 Cablevision Lightpath MOBILE MOST
RECENT [redacted] 2609 AT&T Mobility EMAIL MOST RECENT
[redacted] LANDLINE [redacted]-0646 Verizon
Communications LANDLINE [redacted] 0537 Verizon
Communications LANDLINE [redacted] 0717 Verizon
Communications LANDLINE [redacted]-0221 Verizon
Communications LANDLINE [redacted] 7918 Verizon
Communications LANDLINE [redacted] 7565 Verizon
Communications MOBILE [redacted]-4483 Verizon Wireless

Matched On:
[redacted]

» View Offending Content Source

Add a note about this alert.

```
{
"occurrences":[{"origin":null,"term":"[redacted]"}],"clickablePageLink":true,"screenshot":
```

Alert opened on 08/26/2016 12:20 PM

FIGURE 7.
A sample PII alert generated by the ZeroFOX platform.

6. CONCLUSION

Taken together, the onus is on organizations to protect their enterprise information and formalize some notion of control over social media; a critical gap that organizations spanning different industry verticals have previously left unattended. The vehicle for data exfiltration may have evolved, but the adversary tactics persist from the network, to email, to mobile, to social media. Its simply a matter of recalibrating your defenses to the modern world.

Social media monitoring can actually provide early warning signs to impending threats, allowing organizations to respond by fortifying their defenses. These early warning signs include:

- Detecting social media specific malware and phishing URLs not seen in emails or other delivery mechanisms. Detecting these and refining approaches by re-incorporating them into your protections can help strengthen anti-malware and phishing strategies.

- Impersonators are the basis for many attacks, especially targeted attacks on organizations. These are designed to look almost identical to legitimate accounts, with the exception of perhaps homoglyph characters or a photo-shopped image. Impersonators can be used to establish trust with employees and enable direct messages, which can be used for spear-phishing attacks or custom malware targeted at an employee, allowing for exfiltration of data from the victimized computer.

- Insider threats can be identified by identifying suspicious activities such as an employee frequently dumping data to social media websites like Pastebin.

- Breached data is frequently distributed across the social web, including server IP addresses, logins, passwords, and more. Identifying these instances and changing system passwords prior to intruder access is key to shoring up the organization's defenses.

- Monitoring for PII exposures is a key requirement for many regulatory and industry compliance and important to minimizing fines. Detecting instances of PII can expedite the incident response and close the temporal gap during which data loss occurs.

It's clear that employees are not always aware of their actions and how they may be putting corporate data at risk. This best handled through social media employee education, social media policies (HR and Information Security), and proactive analysis of social media activity. Employees are a key component to deterring social media data loss.

> *"To know your enemy, you must become your enemy"*
>
> -Sun Tzu "The Art of War"

Knowledge and awareness of data loss threats stemming from social media is the first step to defending your data. Since social media resides outside the corporate network, today's perimeter defenses do very little to address DLP. Furthermore, activities are most likely occurring on devices that are outside of your control, through apps that are unmanaged. Understanding many of these outlined threats will better arm you in your monitoring and protection efforts.

7. REFERENCES

1. http://now-static.norton.com/now/en/pu/images/Promotions/2012/cybercrimeReport/2012_Norton_Cybercrime_Report_Master_FINAL_050912.pdf
2. http://www.csoonline.com/article/2978164/advanced-persistent-threats/intel-criminals-getting-better-at-data-exfiltration.html
3. Fraud-and-risk-report-2016 - Callcredit. http://www.callcredit.co.uk/media/2561360/index.htmlhttp://www.fraud.org/component/content/article/2-uncategorised/80
4. http://www.gao.gov/new.items/d08536.pdf
5. http://commerce3.derby.ac.uk/ojs/index.php/sgs/article/download/88/64
6. https://github.com/upgoingstar/datasploit
7. http://www.dailymail.co.uk/news/article-3166496/Jeremy-Hunt-hot-water-breaching-patient-confidentiality-posting-hospital-visit-picture-Twitter-board-names-it.html
8. https://www.propublica.org/article/inappropriate-social-media-posts-by-nursing-home-workers-detailed
9. http://www.telegraph.co.uk/news/uknews/defence/10948490/Troops-leaked-confidential-data-on-Twitter-and-Facebook.html
10. https://news.vice.com/video/selfie-soldiers-russia-checks-in-to-ukraine
11. http://blog.ticketmaster.com.au/ex/protect-yourself-hide-your-barcode-3232
12. http://www.dailymail.co.uk/news/article-3051671/Holidaymakers-post-information-trips-Facebook-face-having-insurance-claims-rejected-home-targeted-burglars-away.html
13. http://www.usmagazine.com/celebrity-news/news/scott-disick-accidentally-pastes-instructions-into-instagram-ad-caption-w207208
14. http://www.bloomberg.com/news/articles/2014-11-25/twitter-cfo-noto-has-an-oops-moment-with-mistaken-tweet
15. http://www.infosecurity-magazine.com/news/check-point-says-social-engineering-attacks-now-a/
16. https://www.consumeraffairs.com/news/private-message-phishing-scam-seeks-to-snag-facebook-users-022615.html
17. https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf
18. http://store.elsevier.com/Metasploit-Toolkit-for-Penetration-Testing-Exploit-Development-and-Vulnerability-Research/David-Maynor/isbn-9780080549255/
19. http://media.kaspersky.com/pdf/kaspersky_lab_ksn_report_the_evolution_of_phishing_attacks_2011-2013.pdf
20. http://barracudalabs.com/wp-content/uploads/2013/06/2011LabsSocialNetworkingStudy.pdf
21. https://www.emc.com/collateral/fraud-report/rsa-online-fraud-report-012014.pdf
22. https://labsblog.f-secure.com/2015/09/03/linkedin-sockpuppets-targeting-security-researchers/
23. https://www.fiverr.com/
24. http://www.csoonline.com/article/2156500/security-awareness/google-facebook-unmask-tech-support-scams.html
25. http://www.darkreading.com/analytics/anatomy-of-a-social-media-attack/a/d-id/1326680
26. https://www2.fireeye.com/rs/fireye/images/rpt-m-trends-2015.pdf
27. http://www.businesswire.com/news/home/20160718005304/en/Ponemon-Institute-External-Cyber-Attacks-Cost-Enterprises
28. http://www.csoonline.com/article/3078338/security/insider-threat-mitigation-techniques-worth-considering.html
29. https://multimedia.3m.com/mws/media/1027626O/study-visual-hacking-experiment-results.pdf
30. http://motherboard.vice.com/read/hackers-stole-over-60-million-dropbox-accounts
31. https://haveibeenpwned.com/
32. http://www.cipd.co.uk/PM/peoplemanagement/b/weblog/archive/2016/09/20/opinion-how-to-use-social-media-to-combat-insider-fraud.aspx
33. http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigation-report_2015_en_xg.pdf

8. TIMELINE REFERENCES

1. https://www.arbornetworks.com/blog/asert/twitter-based-botnet-command-channel/
2. http://www.symantec.com/connect/blogs/trojanwhitewell-what-s-your-bot-facebook-status-today
3. http://blog.unmaskparasites.com/2009/11/11/hackers-use-twitter-api-to-trigger-malicious-scripts/
4. http://www.nartv.org/mirror/shadows-in-the-cloud.pdf
5. https://www.hotforsecurity.com/blog/twitter-controlled-botnet-sdk-at-large-813.html
6. https://www.eff.org/files/2015/02/03/20150117-spiegel-byzantine_hades_-_nsa_research_on_targets_of_chinese_network_exploitation_tools.pdf
7. https://www.skycure.com/blog/linkedout-a-linkedin-privacy-issue/
8. http://informationonsecurity.blogspot.com/2012/11/concealing-data-exfiltration-with.html
9. https://labs.bitdefender.com/wp-content/uploads/downloads/2013/04/MiniDuke_Paper_Final.pdf
10. https://www.f-secure.com/weblog/archives/00002576.html
11. https://www.skyhighnetworks.com/cloud-security-blog/100000-tweets-in-1-day-how-company-discovered-security-breach-using-big-data-analytics/
12. http://www.tripwire.com/state-of-security/incident-detection/hackers-exfiltrating-data-with-video-steganography-via-cloud-video-services/
13. https://www.f-secure.com/weblog/archives/00002764.html
14. http://www.itproportal.com/2015/01/30/cunning-f0xy-new-smart-malware-stealth-trickery/
15. https://www2.fireeye.com/rs/848-DID-242/images/rpt-apt29-hammertoss.pdf
16. https://github.com/DakotaNelson/sneaky-creeper
17. https://newsfromthelab.files.wordpress.com/2015/11/cc-as-a-service.pdf
18. https://github.com/PaulSec/twittor
19. https://securelist.com/blog/research/68838/be2-extraordinary-plugins-siemens-targeting-dev-fails/
20. https://github.com/sensepost/DET
21. https://www.endgame.com/blog/instegogram-leveraging-instagram-c2-image-steganography
22. http://www.bleepingcomputer.com/news/security/the-crylocker-ransomware-communicates-using-udp-and-stores-data-on-imgur-com/