Exploiting Military OpSec through Open-Source Vulnerabilities

Judson C. Dressler[*], Christopher Bronk[†], Daniel S. Wallach[*]

[*]Department of Computer Science
Rice University
{jd25, dwallach}@rice.edu

[†]Department of Information and Logistics Technology
University of Houston
{rcbronk}@uh.edu

***Abstract*—With the ease of use and connectivity provided by social media, there has become a growing tension between military users' personal needs and military operational security. Like everyone, military members post seemingly trivial information and pictures on sites such as Facebook, Twitter, or LinkedIn; all of which can be used, augmented, and aggregated by an adversary to determine the utility and feasibility of possible intelligence targets. In this work, we investigate the current state of DoD policy regarding social media. We then designed an automated approach to determine the amount of openly available information provided by U.S. military members through social media; analyzed it through content analysis; then applied machine learning techniques to learn as much from the provided data as possible. We then ranked the vulnerability of each individual found by the data provided with a simple scoring system; classifying them as least vulnerable, vulnerable, highly vulnerable. In all, we discovered 1100+ potential intelligence targets; hypothesized about potential scenarios in which this publicly available information could have negative consequences; and recommend actions that could mitigate this vulnerability.**

I. Introduction

With the ease of social media, military members and their families can post pictures of an ongoing operation or discuss events of a current deployment from any where at any time. While individuals may feel the information publicly posted is harmless when viewed in isolation, it is precisely the type of information an adversary can leverage for use in intelligence targeting. Bits created in one part of the physical world can easily find themselves elsewhere, allowing an adversary to piece together trivial pieces of information to provide a robust picture of activities, friends, and family; all of which can be used, augmented, and aggregated to create a highly refined list of potential intelligence targets. With this era of greater interconnectedness and easier communication, there is a growing tension between military users' personal needs (keeping in touch with family and community) and military OpSec, or operational security.

Rather than placing a spy in an organization, adversarial nations or terrorist groups can use social media to find valuable targets. Since nearly everyone with an Internet connection has a Facebook, Twitter, LinkedIn or other online profile, adversaries can "spy" on members of the United States military; learn their schedules, habits, interests, discontents, secrets, etc; and then bribe, threaten, or coerce them into turning over sensitive information. While this information used to be private, social networking sites provide an adversary an easy avenue for data collection with little risk or cost.

Our initial hope was that military members would be well aware of the dangers of social media and, while still using these networks, would ensure appropriate security measures were in place. Although Operational Security (OpSec) is frequently discussed during military training, we were able to recognize many questionable posts on Facebook, LinkedIn, and other social networking sites. While a few non-compliant military members is a concern (as one slip could cost lives), we wanted to determine the amount and type of information available about U.S. military members and how widespread the problem might be.

II. Motivation

Based on hundreds of pieces of public and private data ("Big Data"), marketers, financial institutions, and other businesses are creating profiles predicting personal behavior; from how likely we are to buy a product to how likely we are to end up in the hospital [1]. With massive amounts of information freely available or cheaply purchased, what would stop an adversary of the U.S. from profiling its military members and civilian DoD counterparts to develop intelligence prospects? Of course, similar risks extend to anybody (military or civilian, domestic or foreign). This study follows Spang [2], which was done largely by hand, where our work uses automated methods to consider a larger sample. Our goal was to provide the military with thorough actionable information, possible enemy scenarios, and recommended actions to remediate this real-world threat. We expect our guidance will also be applicable well beyond the U.S. military.

III. Department of Defense Guidance

The Department of Defense (DoD) initially attempted to limit the use of open social sites such as Facebook, Flickr, and YouTube due to concerns of security, accountability, and privacy [3]. However, in 2010, the department reversed this ban citing the need for better information sharing and to accommodate younger users who had come to expect social media access, thus allowing use of unclassified .mil computers to access such sites [4]. Whether allowed to use military information systems to access these sites or not, social media and social networking have evolved to become the primary communication method used by today's military members and their families, integrated into all aspects of their lives. With this in mind, a review of the guidance provided by the DoD is necessary.

A. *Operational Security (OpSec)*

The Department of Defense Instruction 5205.02E [5] and its associated manual 5205.02M [6] outline the roles and responsibilities of the OpSec program. It defines the OpSec process as a "systematic method used to identify, control, and protect critical information" and defines critical information as "information that the organization has determined is valuable to an adversary" [5], [6]. While not directly addressing social media sites, the instruction calls attention to the exposure of critical information through aggregation and its mitigation through awareness training and guidance for those "using DoD Internet services, other Internet-based capabilities, emerging technologies, or developing information sharing environments that are accessible across the enterprise." In addition, the manual calls for security to be integrated into new systems as well as procedures to deny adversaries the opportunity to take advantage of publicly available information, especially when aggregated.

B. *Social Media Handbooks*

Following this recognized need for user training, each military service has a social media handbook that derives from DoD Instruction 8550.01 entitled "DoD Internet Services and Internet-based Capabilities" [7]. All of the services see social media as a cheap, effective, and measurable form of communication and even encourage their personnel to use it to share their experiences and to keep it touch when deployed, but personnel must ensure appropriate conduct is maintained between leadership and subordinates. While the majority of each handbook is devoted to the proper methods for official online presences, some tips are provided to military members for unofficial presences.

- Sharing seemingly trivial information online can be dangerous to loved ones and fellow military members, as well as leaving you exposed to identity thieves.
- Do not break OpSec as everything shared online must be considered public.
- Do not geo-tag photos while deployed.
- Differentiate between opinion and official information.
- Be on the lookout for intruders, use appropriate security software, and continually review account and privacy settings.

Notably, Marine Corps provided step by step instructions on securing a Facebook profile; unfortunately, with constant privacy changes made by Facebook, these instructions were quickly out of date. Each service also emphasized that posting personal information such as children's photos, names, schools, ages, and schedules can be dangerous. Each handbook provided guidance on what types of information not to place on social media; however, these documents did not link social media to the danger of becoming a targeted intelligence asset for an adversary [8], [9], [10], [11].

C. *Guidance of Other Nations*

The British and Canadian militaries issue largely similar advice to that given by the United States military, expressing a desire to use social media to help the public understand the challenges the military faces as well as strengthen the bonds between the military member and their family and friends. While they have stressed the need for OpSec within these applications as well as the risk of a simple geo-tagged photo or leaked detail on Twitter, they do not address personal information and how it could be dangerous [12].

The Israeli Defense Forces are much more restrictive. They train their soldiers to not identify themselves or discuss operational issues. They have even instructed their civilians to avoid status updates when rockets or attacks happen nearby, hopefully reducing the possibility of their adversary using social media as a type of battle damage assessment tool [13].

While we were unable to locate Chinese social media guidance, it is important to note that the U.S. (and computer security companies) are profiling Chinese cyber units through the corroboration of digital attack traces with open source data including social media [14].

IV. Research Design

This research utilizes select social media sites and uses content analysis and machine learning algorithms to identify members of the United States military (Army, Navy, Marines, and Air Force) who may, based on their Internet activity, be leaking too much sensitive information, becoming a prime target for an adversary's intelligence activities. We began with automated data collection.

A. *Data Collection*

Facebook is a social utility that connects people to keep up with friends and share photos and videos. We began there because it is the #1 social media site and the second most visited site in the world [15]. On Facebook, each user creates a profile detailing as much or as little personal information as they choose, aside from the mandatory fields of name and age. The data initially entered, as well as photos, videos, and other updates provided by the user, can then be shared throughout their network. Depending on how a user chooses to secure their profile, being a member of a user's network offers access to additional information not shared with non-members. The vulnerability we exploited originates from the failure of Facebook users to appropriately choose security settings preventing the public from viewing potentially personal information [16]. In addition, Facebook's poor default security and constantly changing privacy settings (including many well-publicized fiascos) have left many users with public profiles who may not be aware of exactly how public they are [17].

Using Facebook's application programming interfaces (APIs), we developed a script that searched for users who publicly self-identify as being a member of the United States military and then scraped as much data as possible from their public profile. The script then continued to search, utilizing the identified military member's network of friends, if available. Since the "Employer" category is populated by the user, the field returned varied based on the specific text the user chose to use; for example, members of the U.S. Army may have identified their employer as "US Army", "United States Army", or "U.S. Army". To generate the most comprehensive results possible, we used these and many other variations and merged the results.

All the data we gathered were from public pages. No account intrusions, social engineering, or other deceptive practices were used to gather our data. For better or for worse, Facebook does not verify anything, so users can and will falsely claim to have a military affiliation, present or past. To increase the quality of our results, we manually scrutinized the data and discarded profiles belonging to individuals where inconsistencies were found (i.e., fake sounding name, impossible birthdates, military rank above what is possible for the user's age as age and rank are strongly correlated in the U.S. military, etc). An adversary looking to collect data on U.S. military servicemembers would likely follow a similar approach. *In all, we found 3080 public Facebook profiles across all four military branches.*

Next, we applied this same data collection approach to LinkedIn. LinkedIn is a networking site designed primarily to build one's professional identity online, including discovering professional opportunities and business deals, and getting the latest news and insights into a chosen professional industry [18]. It is the third largest social networking site and #12 on Alexa's most visited sites globally, as well as being the self-proclaimed world's largest professional network [15], [18]. Similar to Facebook, the user controls the amount of information entered as well as the amount of information the public or "connections" can view. In addition, LinkedIn offers an upgraded "Recruiter Corporate" account for only $720 per month for full access to all 300 million LinkedIn members. For this research, we only considered LinkedIn data that is visible to the public, without any special permissions or insider access. We note that any intelligence agency would happily spend the additional money for unrestricted access.

Using LinkedIn's APIs, we developed a script that searched for any military member's name that had been discovered through Facebook. We then manually corroborated these profiles with the information gained through Facebook to ensure it was the same individual, and any additional information gained was added to the target profile. Any inconsistencies found between the Facebook and LinkedIn data resulted in the exclusion of the target profile altogether. *In all, of the 3080 Facebook profiles we found, 902 had corresponding LinkedIn accounts.*

The results from combining the profiles from Facebook and LinkedIn are shown in Tables I and II.

B. Control Group vs. Military

There is no obvious way for us to determine how many military members have social media profiles that are set to less-than-public visibility, or that hide their military affiliation. As a proxy, we decided to consider a "control group" consisting of alumni from two large U.S. public universities. We estimate these two universities to have roughly one million living alumni, yielding a total population of comparable size to the approximately 1.5 million members of the U.S. military. Of course, the age demographics will be quite different between these two populations (i.e., as military personnel age they tend to retire from active duty, while alumni never "retire" from their alma mater), so direct comparisons are only meaningful in broad brushstrokes.

From our control group, we found 3386 public Facebook profiles, versus the 3080 profiles found from U.S. military members. Similarly, from our control group, we found 1217 corresponding public LinkedIn profiles, versus 902 profiles from U.S. military members (see Table I). If we adjust these numbers as percentages of the overall estimated population, we see Facebook and LinkedIn public profiles as 0.21% and 0.06%, respectively, of the military population, and 0.34% and 0.12%, respectively, of the university alumni population.

One can draw a variety of inferences here. Since social network use tends to be more active among younger populations, and the university alumni population skews older than the military population, we can conclude that military personnel are less likely to have public profiles on social networks than the general population. However, these public military profiles seem to share a similar amount of information to the information shared by the university alumni profiles (see, e.g., Table II). From the perspective of military OpSec, it's unacceptable to have *any* such public profiles. An adversary who can convert even one insider can leverage them to a variety of ends. (The recent cases of Bradley Manning and Edward Snowden illustrate how singular individuals may be able to exfiltrate significant volumes of sensitive data.)

Aware of these risks, the U.S. military begins OpSec education in basic training, with a refresher course required annually thereafter. While this training briefly covers social media sites, it only does so when discussing ongoing operations, not personal risk. With the DoD spending $5.29 million on its operational security program in fiscal year 2014 alone [19], and with the seemingly high number of public profiles of military personnel, we argue that the training program in place is inadequate.

C. Scoring System

From the information collected above, patterns of life can easily be determined. A good recruiter would have a plethora

...

TABLE I
TOTAL NUMBER OF FACEBOOK AND LINKEDIN PROFILES DISCOVERED BROKEN OUT BY MILITARY BRANCH

Facebook and LinkedIn Combined						
	Air Force	*Navy*	*Army*	*Marines*	*Total*	*Control*
# of Profiles Found	294	248	267	93	902	1217

...

TABLE II
PERCENTAGE OF COMBINED FACEBOOK AND LINKEDIN PROFILES DISCOVERED CONTAINING THE SPECIFIED INFORMATION

Facebook and LinkedIn Combined						
	Air Force	*Navy*	*Army*	*Marines*	*Total*	*Control*
First	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Last	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Gender	99.0%	98.0%	97.8%	98.9%	98.3%	98.8%
High School	76.5%	77.0%	84.6%	79.6%	79.4%	78.1%
College/University	90.1%	94.4%	87.3%	82.8%	89.7%	100.0%
Graduate School	43.5%	45.6%	26.6%	30.1%	37.7%	41.6%
Personal Photo	77.2%	81.1%	79.8%	86.0%	79.9%	81.1%
Age/DOB	60.2%	59.3%	60.3%	63.4%	60.3%	59.2%
City	99.3%	96.8%	97.4%	96.8%	97.8%	98.4%
Marital Status	34.7%	31.1%	30.3%	23.7%	31.3%	39.9%
Current Place of Work	100.0%	100.0%	100.0%	100.0%	100.0%	95.7%
Spouse Name	9.5%	12.1%	9.7%	22.6%	11.6%	9.3%
Anniversary	3.1%	4.0%	3.8%	5.4%	3.8%	4.6%
Phone #	0.7%	0.4%	0.8%	3.2%	0.9%	1.3%
Home Address	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
Interests	73.1%	77.8%	77.5%	68.8%	75.3%	67.6%
Professional Skills	59.9%	62.9%	54.3%	58.1%	58.9%	66.4%
Viewable Network	79.3%	75.8%	74.9%	77.4%	76.8%	79.3%
Rank	4.8%	7.3%	5.2%	7.5%	5.9%	0.1%
Viewable Timeline	64.6%	75.0%	51.3%	48.4%	61.9%	74.9%
Geotagged Photos	34.0%	41.1%	48.7%	47.3%	41.7%	48.2%
E-mail address	6.5%	9.3%	9.7%	8.6%	8.4%	11.0%
Work Experience	77.6%	84.3%	83.2%	79.6%	81.3%	87.3%
Skills	49.7%	55.7%	64.8%	51.6%	56.0%	65.2%
Clearance	17.4%	19.4%	27.0%	20.4%	21.1%	2.0%

of this Internet-originated information to draw from to create and foster a fake relationship. With the provided information above, an intelligence officer may easily create a short list of preferred candidates. We used two factors to determine the vulnerability of a United States military member to a foreign intelligence service: 1) the access an adversary has to an individual; 2) the individual's access to classified information.

For each target profile identified in the data collection phase, the following pieces of information, if provided, were recorded as a measure of "Access to Individual" by an adversary: first and last name, schools attended, home address, photos, contact phone, e-mail address, gender, anniversary, rank, professional skills and interests, and viewable network and timeline. For "Access to Information", the following pieces of information were recorded and measured: holding a secret or top secret clearance, holding a management position, holding a position allowing for access to information, and employment with the United States military.

Since not all information is of equal significance to an adversary, we created a weighted scale. We assigned each category a value between 1 and 5 based on the relative importance of the information, with 5 meaning the information would be most valuable to an adversary and 1 meaning the information is least valuable. Our weightings are shown in Tables III and IV. This scoring system is a slight adaptation of the one proposed in Spang [2].

The summation of the numeric values for each critical piece of information discovered provided a score for each of the two factors. The combination of the two factors was used to determine the vulnerability of the U.S. military member for exploitation by an adversary (i.e., the higher each factor, the more vulnerable a member was).

Figure 1 is a 2-dimensional scatter plot of the numeric scoring results for "Access to Information" vs "Access to Individual" from the combined Facebook and LinkedIn data. The more personal information a member provided contributed to a higher score on the X-axis while more information about a member's access to classified or critical information contributed to a higher score on the Y-axis.

For the "Access to Information" factor, a sufficient baseline is if the member has a security clearance, as represented by the horizontal line at an "access to information" score of 5 in Figure 1. This allows the adversary to know the individual has access to restricted areas and classified systems.

TABLE III
NUMERICAL-BASED SCORING SYSTEM AND RATIONALE FOR EACH CATEGORY OF ACCESS TO INDIVIDUAL DISCOVERED

Access to Individual		
Critical Information	*Assigned Value*	*Rationale*
Home Address	5	Provides direct physical access to individual, probably income
Geotagged Photos	1 - 5	Can provide direct physical access to individuals. 1: less than 5 photos; 2: less than or equal to 15 photos; 3: greater than 15 photos; 4: greater than 15 photos with one location on several occasions; 5: 15 photos and more than 1 location on several occasions
Spouse Name	4	Spouse allows for easy identification; opens another avenue for intelligence; solution to many challenge questions (met them, married them, date married), also allows easier manipulation of primary target
Contact Phone	3	Challenge question answer; provides another avenue for social engineering; Opens potential target of individuals phone
Self Photo	3	Makes target easily identifiable for espionage; also may provide insights into interests/family
Viewable Timeline	1 - 3	Can provide information including location, interests, professional skills, friend network, phone number, birthdate, marital status, anniversary, education. Scored on a sliding scale from 1-3 to differentiate between a public wall that's largely empty versus somebody who posts everyday versus somebody who posts personal information (complaining about ex-wife, dislike for job, unhappiness with deployments, etc).
Professional Skills	3	Allows for the specific targeting of individuals with access to specific data/technology
Age/DOB	2 - 3	Often used for pin numbers, allows for personal identification, often a challenge question. 2 points for age, 3 points for date of birth
Last Name	2	Readily available, easily leads to other data but must be combined with other data to be effective
High school	2	May provide solution to online challenge questions; valuable for social engineering/phishing schemes
College	2	Provides details for social engineer/phishing schemes; also allows determination of possible current job responsibilities
Graduate School	2	Provides details for social engineer/phishing schemes; also allows determination of possible current job responsibilities/level of responsibility
Interests	2	Challenge question answer; also valuable for social engineering/phishing/water hole schemes
Viewable Network	2	Social network allows for larger surface area to find easiest target, as well as providing adversary legitimacy to come after primary target
Anniversary	2	Valuable for social engineering/phishing schemes; also challenge questions
Rank	2	Indication of amount of responsibility, age, salary; can be used during promotion cycles for social engineering/phishing
E-mail address	2	Used as account/user name on many sites, provides avenue for social engineering/phishing
First name	1	Readily available, easily used to lead to other data but must be combined with other data to be effective
Gender	1	Most of the time easily determinable by first name, not necessarily helpful to enemy
Marital Status	1	Easily found in public records, can lead to further investigation on spouse/family
Current City	1	Narrows field of potential people to coerce; can be combined with other data to find home address/tax records for specific target

TABLE IV
NUMERICAL-BASED SCORING SYSTEM AND RATIONALE FOR EACH CATEGORY OF ACCESS TO INFORMATION DISCOVERED

Access to Information		
Information Category	*Assigned Value*	*Rationale*
Holding a Secret to Top Secret Clearance	5	Clearly states individual has access to classified materials
Holds a position allowing for access to information	2 - 3	Indicates individual probably has access to classified materials. If management position, 3 points, otherwise 2. To determine if management position was held, key terms were searched for in job title, such as commander, director, etc. as well as officer ranks O-3 or higher and enlisted ranks E-7 or higher.
Employed by U.S. Military	1	Indicates individual is military member

On the "Access to Individual" side, the cause for concern has always been if the person can be identified from the data collected. While none of this data is considered personally identifiable information (PII), technology and the wide availability of information about people enables the aggregation of various pieces of non-PII to produce PII. Sweeney [20] found that the combination of zip code (most of the time easily discerned from city), birth date, and gender was sufficient to uniquely identify 87% of individuals in the United States. With roughly 7.2 billion people in the world, it takes 33 bits of entropy to identify an individual [21]. Gender, for example, is worth one bit; reducing the worlds population by roughly half. However, the entropy decreased by other pieces of information is not as easily discernible: living in a heavily populated city or zip code would not be worth as many bits as a sparsely populated city or zip code.

Given our weighting scheme, we can declare arbitrary cutoffs above which an individual moves into different categories (i.e., least vulnerable, vulnerable, highly vulnerable). By studying the various criteria and weights that we assigned to them, we considered an "access to information" score about 5 to be a significant threshold. Similarly, we felt that "access to individual" scores above 15 and above 25 seem to be significant differentiators among the people in our dataset. The combination of these thresholds and the assignment to risk categories is shown in Figure 1. With these thresholds, *184 military members are vulnerable and 40 are highly vulnerable.*

We note that absent information from Facebook and

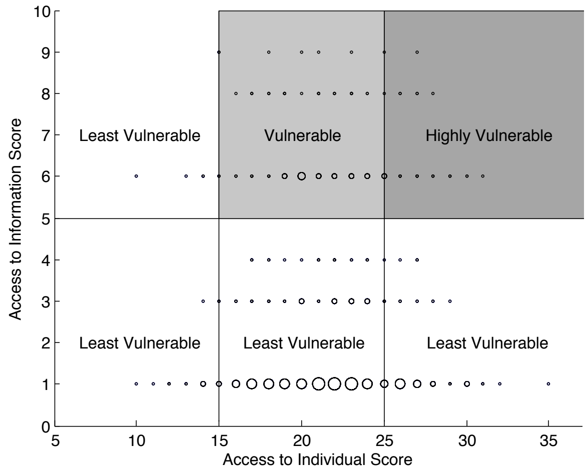

Fig. 1. Vulnerability of Military Members via Social Media. Four different sized circles were used corresponding to specific ranges of users within those buckets (larger circle = more users, smaller circle = less users).

LinkedIn might be reconstructed from other sources. Home addresses, for example, can be determined from property records, telephone records, voter registration databases, and a variety of other public sources. Spousal names may be discovered from marriage records. Age, date of birth, and city of birth will appear in birth records. Education records may be public as well (e.g., many unversities publish undergraduate theses online, which will have cover pages indicating author and year; doctoral students produce publications which will also contain contact information). An intelligence analyst would certainly use these other sources to individually evaluate each potential asset. While we did not pursue such a manually-intensive analysis, we wanted to consider how much missing data could be predicted automatically.

D. Predict Missing Values

Since social media users may omit information about themselves, either out of neglect or concerns for privacy, discovering this additional information makes the profile more of an adversarial intelligence target. Therefore, we wish to apply machine learning algorithms on the information already provided by the user and their network in an attempt to predict the omitted information. To that end, we trained a classifier based on the example of profiles that contained the desired information (label) so it can be applied to the unlabeled profiles to predict labels for them. To do this, we created a graph representation of the Facebook and LinkedIn social networks collected, $G(V,E,W)$ where:

Nodes V: The set of nodes representing the user profiles collected from Facebook and LinkedIn.

Edges E: An edge $(i,j) \in E$ between two nodes v_i, v_j represents a “friendship” or “connection”.

Edge Weights W: The weight w_{ij} on an edge between nodes v_i, v_j indicate the strength of similarities between the two nodes.

Note: Since friendships/connections in Facebook/LinkedIn are reciprocal, the edges in graph G are undirected.

Problem Statement: Given the graph $G(V,E,W)$ with a subset of nodes $V_l \subset V$ labeled and $V_u = V/V_l$ the set of unlabeled nodes (i.e., the information in V not contained in V_l). Let Y be the set of m possible labels, and $Y_l = y_1, y_2, \ldots, y_l$ be the initial labels on nodes in the set V_l. The goal is to infer labels Y on all nodes V of the graph G.

The labels from neighboring nodes are used as the primary source because links between nodes in social networks indicate a relationship between the individuals that the nodes represent. In particular, a link indicates a higher probability of similarity between the linked individuals because friendships usually occur between individuals of similar nature and interests.

Therefore to derive the weights (W) for graph G, the following factors were considered between two nodes i and j: the number of friends i and j have in common, the number of known labels shared by i and j, and how often i and j communicate (i.e., show up in each others' Facebook timeline). Each factor was normalized by its maximum value in G and weighted equally.

To prepare the data for the machine learning algorithms, categorical labels (such as high school, college, graduate school, rank, and current city) were made into individual features using a binary flag for each category. This supervised data set was then used with a variety of standard machine learning algorithms: Naïve Bayes (NB), 3-Nearest Neighbors (3NN), Support Vector Machine (SVM), and Random Forest (RF).

Naïve Bayes (NB) Classification: A standard Naïve Bayes model which for each user-label pair predicts the probability that the label belongs to that user.

3-Nearest Neighbors (3NN): A standard k-NN algorithm is used in which the user's label in each category is classified by a majority vote of its three closest (highest weighted) neighbors.

Naïve Bayes and 3-Nearest Neighbors algorithms were chosen based on the social phenomena of homophily, which states that individuals tend to associate with individuals similar in nature.

Support Vector Machine (SVM): An SVM classification model was built for each label. The Radial Basis Function (RBF) kernel was used and for categorical labels, a one-versus-many approach was employed. SVMs perform well on data sets with many attributes (labels), even with very few data points on which to train the model. It also has strong regularization properties (less prone to overfitting) which allow it to generalize to new data easily.

Random Forest (RF): Random Forest is an ensemble learning method (very closely related to nearest neighbor) for classification that constructs decision trees at training time and outputs the mode of the classes output by the individual trees. Put simplistically, it randomly produces trees of weak learners who work together to form a strong learner. The Random Forest algorithm also does not overfit the data and allows for ranking variable importance in the classification.

All our experimentation was performed in a 10-fold cross validation setting, to ensure each model would generalize to an independent data set. To begin each test, the data set of only the profiles where we know the data point were divided into 10 subsamples of equal size, 9 of the subsamples were used as the training set, while 1 was used as the test (or validation) set. The process was repeated 10 times, with each of the 10 subsamples used exactly once as the validation set. In addition, the classification process was made iterative; in other words, the classification process spreads information to new places in the graph which is then fed into the features used for the next round of classification. The results from each of the 10 runs was then averaged to produce a single estimation. For all tasks, test set classification rate (true positive) was reported (as seen in Table V). As an example, the 3NN classifier had a 89.18% test set prediction accuracy for gender. This means that after being trained, this classifier was able to predict the correct gender of the profiles in the test set 89.18% of the time.

...

TABLE V
TEST SET PREDICTION ACCURACY FOR NAÏVE BAYES, 3-NEAREST NEIGHBOR, SUPPORT VECTOR MACHINE, AND RANDOM FOREST CLASSIFIERS FOR EACH INFORMATION CATEGORY

	Algorithm			
Information Category	*NB*	*3NN*	*SVM*	*RF*
Age	37.50%	45.96%	34.19%	41.54%
Age (+/- 1)	88.05%	86.58%	81.43%	83.46%
High School	51.26%	50.14%	47.35%	52.37%
College	76.14%	78.99%	74.41%	81.58%
Graduate School	33.82%	43.82%	29.12%	36.18%
Interests	88.51%	92.49%	89.84%	91.90%
Rank	90.57%	88.68%	88.68%	90.57%
Gender	87.94%	89.18%	85.01%	88.50%
Marital Status	78.37%	81.56%	74.11%	77.66%
Current City	65.99%	72.11%	61.68%	69.95%

Intelligence analysis is based on contingent, not absolute, prediction; gauging the likely outcomes based on the range of possible scenarios [22]. From an attacker's perspective, a classifier with 80% or better accuracy is more than sufficient, as perfect information is neither necessary nor ever available. Thus, with these algorithms, additional information such as approximate age, college, interests, rank, gender, and marital status could be inferred for individuals who did not provide all of the details themselves. Combining the profiles from Facebook and LinkedIn with the predictions of the best machine learning algorithms for each category (of those achieving an 80% or better accuracy), the results are shown in Table VI.

As seen in Table V, the classifiers with the best results rely on the person's closest friends in predicting the value of the missing information. It seems, with little surprise, that individuals are closest with friends of the same age, background, interests, rank, and marital status (i.e. Academy graduates tend to still be close friends with other Academy graduates and married individuals tend to associate more with married individuals). *With the additional information from the machine learning classifiers, 57 more members moved into the highly vulnerable category, 4 into the vulnerable category.* If we had a ground truth to create a classifier for the clearance category, our results would have been much higher as approximately 79% of the profiles collected were immediately deemed not vulnerable due to this constraint.

...

TABLE VI
PERCENTAGE OF PROFILES DISCOVERED CONTAINING THE SPECIFIED INFORMATION WHEN FACEBOOK, LINKEDIN AND BEST MACHINE LEARNING ALGORITHM RESULTS WERE COMBINED

Updated Profiles on Both Facebook and LinkedIn Using Best Machine Learning Classifier			
	Before	*After*	*Difference*
Age (+/-1)	60.31%	95.23%	34.92%
College	89.69%	98.00%	8.31%
Interests	75.28%	98.12%	22.84%
Rank	5.88%	91.02%	85.14%
Gender	98.34%	99.78%	1.44%
Marital Status	31.26%	87.25%	55.99%

E. Finding Other Military Members

For the final aspect of this research, our goal was to increase the number of possible targets found through social media. Again, machine learning algorithms were used in an attempt to recognize individuals who do not self-identify as a member of the military but who actually are. To do this, the same graph setup as before was used, however, this time it was limited to members of the U.S. Air Force and their associated connections; as we were able to verify current employment through an Air Force internal human resources database. A classifier was trained based on the example of profiles that contained current employment so it can be applied to the unlabeled profiles in order to predict whether or not they were currently employed by the Air Force. Using only the publicly available social network connections from the 294 self-identified Air Force employees who had both a Facebook and LinkedIn profile, 87 new targets were discovered. Of these, 74 turned out to be current Air Force members. (For this study, the internal Air Force database was not used for any other purpose beyond testing false-positive rate.) Overall, this study achieved an accuracy of 85.3% in predicting military status with a false positive rate of 14.7%. Scoring these profiles with the system described in Tables III and IV, we determined 13 of the 74 were vulnerable, 4 highly vulnerable.

If these same ratios held for the other three military branches, this search of publicly available social profiles on Facebook and LinkedIn, combined with machine learning algorithms, would have *resulted in a total of 1168 potential intelligence targets, of which 223 are vulnerable, with 109 highly vulnerable.*

V. SCENARIOS

A few hypothetical scenarios are presented in order to emphasize the negative consequences that could arise from the massive amounts of data easily gathered about our service members.

Tagging Military Targets While Traveling: Intelligence adversaries use this database gleamed from social media to immediately identify a U.S. military member, while in civilian clothes, on vacation in a foreign country (using additional information from hotel reservation, credit cards, passport, etc). The member is then captured and held for intelligence or ransom. This may be exceptionally dangerous as military members frequently travel overseas on civilian carriers.

Faux Employment: A military officer has complained multiple times on Facebook account about financial problems and has begun looking for outside work income using a LinkedIn account. Knowing the officer's background and access, an adversary uses a consulting firm as a guise to engage with the officer and gather information.

Prisoner of War / Kidnapping: An officer is captured while in active duty combat. According to Article 5 of Military Code of Conduct, the officer should only disclose name, rank, serial number, and date of birth to the captors. However, through social media, the captors have access to family names and pictures as well as social network commentary (e.g., about the current war, or about broader political opinions), allowing them to manipulate their captive as well as to broadcast more effective propaganda.

Watering Hole/Phishing Attacks: With cyber espionage being a convenient and powerful option for many adversaries, they will naturally employ it to gain insight into U.S. capabilities. Through social media, they may infer a soldier's rank, base, and operational assignment. Knowing that soldier's friends, school history, and current interests, they are fully armed to mount a targeted email attack (i.e., spear phishing) or even an in-person attack (i.e., watering hole) against the officer. Kaspersky Labs assesses this to be a prominent attack vector. Over 20% of the 37.3 million phishing attacks in 2012-2013 occurred through social media [23].

VI. Recommended Actions

Social media enables immense capabilities for the military through easy dissemination of information and increasing esprit-de-corps. This paper does not refute this nor does it recommend going back to a world without social media, as personal and mobile use of these social platforms is immense; banning their use would not necessarily solve these problems. However, certain risk mitigation strategies must be in place to protect the personal data of our military members as well as the security of our military operations. We present a range of options, recognizing that not all of them will be palatable.

As it stands right now, users are the weakest link, as they are inadvertently divulging personal (and possibly sensitive) information through social media. Few technical controls can defend against clever social engineering attacks, whether phishing, faux profiles, or watering holes. Therefore, the threat caused by social media must be made aware to military members through periodic training of policy, guidance, and best practices; a start would be its inclusion, or stronger emphasis, in the annual information assurance and OpSec training. This training should include use cases of the dangers posed. Members should also be advised to secure and possibly scrub their profiles prior to deployments.

Another option would be integrating social media into the security clearance investigative process; notably, there are no social media questions in the current paperwork. Even very basic questions (e.g., "do you have a Facebook profile? Is it visible to the public?"), perhaps integrated with the efforts of the agents conducting the clearance investigations, would help military users of these services to realize when they've inadvertently made their public profiles too revealing.

Furthermore, an obvious approach would be for the military to continuously conduct "opposition research" on itself through the methodology described in this paper. A continuous process like this would mean that "friendlies" would hopefully discover these vulnerabilities and move to correct them prior to adversaries exercising their own opportunities. This might also be combined with "red team" exercises in a variety of forms (e.g., a red team might use social media as part of a social engineering effort to gain unauthorized access).

There's also a role for regulatory action, which might restrict the "default" public visibility of user profiles, or might constrain the ability for third-party brokers to buy, sell, and aggregate such data without user consent. While such data aggregation services might be primarily of use to Internet advertisers, intelligence agencies might also be able to benefit from the low cost data flows in these environments. (A full study of the degree to which cookies and other advertising identifiers can be used to develop profiles on military users would be an excellent area for future research.)

VII. Conclusion

The Internet was designed with the concepts of reliability and free flow of data, linking all corners of the globe together using common protocols. In wiring together the planet, it has also provided U.S. adversaries a fast, adaptable operating environment that provides a significant cost-benefit advantage relative to traditional targeting techniques. We have shown how a foreign intelligence service or non-state actor can easily discover United States military members through simple open-source querying of social media. While our methodology was limited to passive observation of publicly available information and machine learning algorithms, it produced a rich target set.

Through open source Facebook and LinkedIn data, we found 184 vulnerable military members, with 40 classified as highly vulnerable. Further, we expanded these numbers by inferring omitted profile details through machine learning techniques, resulting in a total of 1168 potential intelligence targets, of which 223 are vulnerable, with 109 highly vulnerable. Further, unstructured open-source research or merging the results with information purchased from a data broker could have further boosted these results. We find, consistent with Spang [2], that, *military members provide the type of personal and professional information in their social network that adversaries spend years and significant resources attempting to develop.*

It is important to note that the digital corpus of information holding potential analytic value is growing. According to an International Data Corporation study in 2012, 25% of the digital universe contained information that might be valuable if analyzed, with only 0.5% currently being analyzed [24], and civilian data mining firms are certainly working on increasing that number. Data are being collected about when individuals are home, their heart rate during a run, or how well they are sleeping at night. The full implications of accumulating health, browsing history, purchasing habits, social behaviors, religious and political affiliations, and finances are not well understood. This data aggregation will have important privacy ramifications for civilians, and even more so for military personnel. This makes it essential for the military to stay ahead of its potential use to target military personnel.

This study focused on the availability of open source data in identifying and prioritizing members of the United States Armed Forces from a counter-intelligence perspective. However, recent threats from terrorist organizations bring to light another disturbing revelation of this data's potential use. A July 2014 Law Enforcement Bulletin warned of a "continued call - by Western fighters in Syria and terrorist organizations - for lone offender attacks against U.S. military facilities and personnel" followed by a call by Islamic State militants to "scour social media for addresses of [military] family members and to show up and slaughter them" [25]. This call has been fulfilled multiple times since March 2015 with ISIS posting a 'kill list' including the names and addresses of hundreds of U.S. military personnel online along with the message "kill them in their own lands, behead them in their own homes, stab them to death as they walk their streets thinking that they are safe" [26]. U.S. military personnel must remain vigilant and work to better secure their privacy.

References

[1] M. Woodruff. (2014, Apr.) The secret way companies are using big data to score you. [Online]. Available: http://finance.yahoo.com/news/the-secret-way-companies-are-using-big-data-to-score-you-135018683.html

[2] J. C. Spang, "Open source information, social networking, and IC employees: An analytic vulnerability assessment," Master's thesis, National Defense Intelligence College, 2011.

[3] M. Drapeau and L. Wells, *Social Software and National Security: An Initial Net Assessment*, 2009.

[4] *Directive-Type Memorandum 09-026 Responsible and Effective Use of Internet-Based Capabilities*, Department of Defense, Feb. 2010. [Online]. Available: http://www.defense.gov/NEWS/DTM2009-026.pdf

[5] *DoD Operations Security (OPSEC) Program Instruction 5205-02E*, Department of Defense, 2008. [Online]. Available: http://www.dtic.mil/whs/directives/corres/pdf/520502e.pdf

[6] *DoD Operations Security (OPSEC) Program Manual 5205.02-M*, Department of Defense, 2008. [Online]. Available: http://www.dtic.mil/whs/directives/corres/pdf/520502m.pdf

[7] *DoDI 8550.01 DoD Internet Services and Internet-Based Capabilities*, Department of Defense, 2012. [Online]. Available: http://www.dtic.mil/whs/directives/corres/pdf/855001p.pdf

[8] *The United States Army Social Media Handbook Version 3.1*, US Army Office of the Chief of Public Affairs, 2013. [Online]. Available: http://www.slideshare.net/USArmySocialMedia/army-social-media-handbook-2012

[9] *The Social Corp, The U.S.M.C. Social Media Principles*, USMC Division of Public Affairs, 2011. [Online]. Available: http://www.jbsa.af.mil/shared/media/document/AFD-120412-038.pdf

[10] *Air Force Social Media Guide*, Air Force Public Affairs Agency, 2013. [Online]. Available: http://www.af.mil/Portals/1/documents/SocialMediaGuide2013.pdf

[11] R. Mabus, *Internet-Based Capabilities Guidance: Unofficial Internet Posts*, 2010. [Online]. Available: http://www.public.navy.mil/bupers-npc/reference/messages/Documents/ALNAVS/ALN2010/ALN10057.txt

[12] M. Piesing, "Tweeting the Taliban: Social media's role in 21st century propaganda," *Wired*, 2012. [Online]. Available: http://www.wired.co.uk/news/archive/2012-03/20/military-social-media

[13] N. Ungerleider, "Inside the Israeli military's social media squad," *Fast Company*, 2012. [Online]. Available: http://www.fastcompany.com/3003305/inside-israeli-militarys-social-media-squad

[14] N. Perlroth, "2nd China Army unit implicated in online spying," *The New York Times*, 2014. [Online]. Available: http://www.nytimes.com/2014/06/10/technology/private-report-further-details-chinese-cyberattacks.html?_r=0

[15] *The Top 500 Sites on the Web*, Alexa, Jul. 2014. [Online]. Available: http://www.alex.com/topsites

[16] *Privacy*, Facebook, Jul. 2014. [Online]. Available: http://www.facebook.com/help/445588775451827

[17] D. Goodin, "Facebook page very much public, even when set as private," *The Register*, 2010. [Online]. Available: http://www.theregister.co.uk/2010/10/25/facebook_privacy_bypass/

[18] *What is LinkedIn?*, LinkedIn, Jul. 2014. [Online]. Available: http://www.linkedin.com/static?key=what_is_linkedin

[19] *RDT&E Budget Item Justification for Defense Operations Security Initiative*, Office of Secretary of Defense, Mar. 2014. [Online]. Available: http://www.stratvocate.com/files/osd-p707/osd.html

[20] L. Sweeney, *Simple Demographics Often Identify People Uniquely*, 2000.

[21] Worldometers. (2014, Oct.) World population. [Online]. Available: http://www.worldometers.info/world-population

[22] L. K. Johnson, *Strategic Intelligence*. Greenwood Publishing Group, 2007.

[23] *The Evolution of Phishing Attack: 2011-2013*, Kaspersky Labs, 2013. [Online]. Available: http://media.kaspersky.com/pdf/Kaspersky_Lab_KSN_report_The-Evolution-of-Phishing-Attacks-2011-2013.pdf

[24] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *International Data Corporation*, 2012. [Online]. Available: http://www.emc.com/leadership/digital-universe/2012iview/index.htm

[25] J. Winter. (2014, Sep.) Law enforcement bulletin warned of ISIS urging jihad aattack of us soil. Fox News. [Online]. Available: http://www.foxnews.com/world/2014/09/17/law-enforcement-bulletin-warned-isis-urging-jihad-attacks-on-us-soil/

[26] O. Darcy. (2015, Mar.) 'islamic state hacking division' posts kill list with purported addresses of u.s. military members. [Online]. Available: http://www.theblaze.com