Privacy Risk Assessment on Online Photos

Haitao Xu[1,2(✉)], Haining Wang[1], and Angelos Stavrou[3]

[1] University of Delaware, Newark, DE 19716, USA
{hxu,hnw}@udel.edu
[2] College of William and Mary, Williamsburg, VA 23187, USA
[3] George Mason University, Fairfax, VA 22030, USA
astavrou@gmu.edu

Abstract. With the rising popularity of cameras and people's increasing desire to share photos, an overwhelming number of photos have been posted all over the Web. A digital photo usually contains much information in its metadata. Once published online, a photo could disclose much more information beyond what is visually depicted in the photo and what the owner expects to share. The metadata contained in digital photos could pose significant privacy threats to their owners. Our work aims to raise public awareness of privacy risks resulting from sharing photos online and subsequent photo handling conducted by contemporary media sites. To this end, we investigated the prevalence of metadata information among digital photos and assessed the potential privacy risks arising from the metadata information. We also studied the policies adopted by online media sites on handling the metadata information embedded in the photos they host. We examined nearly 100,000 photos collected from over 600 top-ranked websites in seven categories and found that the photo handling policy adopted by a site largely varies depending on the category of the site. We demonstrated that some trivial looking metadata information suffices to mount real-world attacks against photo owners.

1 Introduction

With the proliferation of cameras, especially smartphone cameras, it is now very convenient for people to take photos whenever and wherever possible. Furthermore, the prevalence of online social networks and photo-sharing sites greatly facilitates people to share their digital photos with friends online. Every day, around 1.6 million photos are shared on Flickr [1], one of the largest online photo sharing sites. In their rush to share digital photos online, well-intentioned Internet users unwittingly expose much hidden metadata information contained in the digital photos. The metadata information such as camera serial number may seem relatively innocent and trivial but could pose privacy threats to photographers[1] and the people depicted in the photo. Unfortunately, one study [14] shows that up to 40 % of high-degree participants do not even know the term

[1] By photographer we mean the person who took the photo rather than who works as a professional photographer.

H. Bos et al. (Eds.): RAID 2015, LNCS 9404, pp. 427–447, 2015.
DOI: 10.1007/978-3-319-26362-5_20

metadata. The situation becomes worse concerning the fact that a photo could linger on the Web for many years.

During the spread of a digital photo, online social network (OSN) services and other media sites usually serve as the sink. Online media sites often compress and resize the photos they host for space saving. For instance, Instagram uses a resolution of 640*640 pixels for all its photos and automatically resizes any larger photos. Media sites may even remove the metadata information in their hosted photos. However, users usually do not know what online services will do with their uploaded photos [14]. Thus, it is important to raise public awareness of the potential privacy risks posed by metadata leakage and increase their knowledge of how online media sites handle the photos they upload.

Based on the life cycle and the propagation process, we create a taxonomy to classify digital photos into three different stages: "fresh," "intact," and "wild." "Fresh" photos are just freshly taken with a camera. "Intact" photos have been uploaded online but remain intact from the hosting sites. "Wild" photos may have been post-processed multiple times by the hosting sites. In this paper, we perform a data-driven assessment of privacy risks on contemporary digital photos. Specifically, we examine digital photos at the three stages in terms of metadata information contained and potential privacy risks, and we further explore the photo handling policies adopted by online media sites.

To obtain a representative dataset for our study, we collected nearly 200,000 photos in total in various ways including soliciting freshly taken photos through crowdsourcing, downloading original sized, intact photos from a major photo sharing site, and crawling "wild" photos from Google Images and over 600 top ranked websites. We examined the metadata information embedded in these photos and found that metadata was prevalent among photos at each of the three stages. We paid special attention to the metadata fields that may give rise to great privacy concerns. We found that about 10 % of "fresh" photos were tagged with GPS coordinates while 27 %–37 % of "intact" photos and only about 1 % of "wild" photos contained GPS information. We also measured the percentages of photos containing other sensitive metadata information including a photographer's name and modification history.

To understand how a photo is processed after being shared online, we also investigated online sites' policies on handling photos based on 97,664 photos crawled from 679 unique top sites in seven categories—"social networking," "news," "weblog," "college," "government," "shopping," and "classified"[2] sites. We found that photo handling policies adopted by online sites vary with different categories. The "college" and "government" sites hardly resize the photos they host or remove the embedded metadata information. However, the sites in the other categories are more likely to resize the photos and remove the metadata information.

In addition to the sensitive metadata information embedded in a photo, we demonstrated that some other trivial looking metadata information could be exploited to launch re-identification attacks against photo owners. For 62.6 % of

[2] "Classified" refers to the classified advertisements sites such as Craigslist.

Table 1. List of metadata information typically included in a digital photo.

Category	Information	Fields
When	Date Time	Create time, modify time
Where	Location	GPS coordinates, city/state/country
How	Device Info.	Camera make, model, serial number, light source, exposure mode, flash, aperture settings, ISO setting, shutter speed, focal length, color information
Who	People	Artist's name
What	Description	Title, headline, caption, by-line, keywords, copyright, special instructions
Modification	Modification History	Create tool, xmp toolkit, history action, history when, history software agent, history parameters

unique photographers, we were able to uncover their both online and real-world identities with just one photo they ever took and posted online.

The remainder of the paper is organized as follows. We provide background knowledge in Sect. 2. We describe data collection methods for "fresh" photos and characterize them in Sect. 3. We examine "intact" photos in Sect. 4. We characterize "wild" photos and investigate online sites' photo handling policies in Sect. 5. We demonstrate the re-identification attack in Sect. 6. We discuss the limitation of this work and propose our future work in Sect. 7. We survey the related work in Sect. 8 and conclude the paper in Sect. 9.

2 Background

In this section, we first give an overview of the metadata information typically contained in a digital photo, then discuss the potential privacy concerns, and finally illustrate the three stages we define for digital photos.

2.1 Metadata Information in a Photo

There are three most commonly used metadata standards for photos: EXIF, XMP, and IPTC. They often coexist in a photo and constitute the main part of the photo metadata. Table 1 lists the metadata fields typically included in a photo grouped by category.

A digital photo typically contains ample metadata information. When a shot is taken, the camera automatically embeds into the photo all the information it knows about the camera itself and the photo. In addition, users can add their own descriptive information with image processing software. Specifically, typical metadata information can be summarized as follows: (1) *when* – when the photo is created and modified if applicable, (2) *where* – the exact location (GPS coordinates and altitude) at which the photo is captured if a GPS receiver is equipped

and enabled, or coarse-grained location information such as city/state/country, (3) *how* – the camera device used, its make, model, serial number, light circumstances (sunny or cloudy, flash on or off), exposure (auto or manual), and all other parameters used, (4) *who* – the photographer and the people depicted in the photo if manually added during post processing, (5) *what* – title, headline, caption, keywords, copyright restriction, and other detailed descriptions added for logging, organization or copyright protection, and (6) *modification* – if the photo is modified, on what date and time, by what software on what computer, and the specific actions done to the photo.

2.2 Potential Privacy Concerns Arising from Photo Metadata

Most metadata fields may look innocent and trivial. However, some could raise serious privacy concerns. We highlight several sensitive metadata fields below.

Geolocation. Contemporary cameras and smartphones are typically equipped with GPS functions. When taking photos with these GPS-enabled devices, geolocation information is automatically saved into the metadata. For a photo posted online, anybody able to access it could check the metadata information and may get the geolocation where the photo was taken. This definitely violates the privacy of the photographer and the people depicted. For instance, the time and location embedded in an online photo indicated that a public figure had been at an embarrassing location and not where he claimed to have been [5]. Moreover, a geo-located photo obviously taken at home and depicting high-value goods may give burglars incentives. In addition, young parents usually like to post many photos of their kids online, which may raise great concerns because the photos tagged with GPS coordinates could disclose the exact locations of where their kids live, play, or study.

Photographer's/Owner's Information. Some photos explicitly contain in the metadata the photographers' information, among which the name information is most commonly seen. No matter whether such information is embedded with or without the photographers' awareness, disclosing such information may cause identity leakage, especially given the availability of geolocation information in the metadata.

Modification History. When post processing a digital photo, an image processing software like Adobe Photoshop and Apple iPhoto often automatically embeds into the photo the detailed modification information, represented by three metadata fields: History When, History Software, and History Parameters. Table 2 presents an example of the embedded modification information in a photo. For the convenience of illustration, we add the photo's shot time in the table. It clearly shows that the photo has been processed twice in less than one month since it was taken on July 16, 2014. And two versions of Adobe Photoshop on one or two Macintosh computers were ever used for format conversion and save actions.

A photographer may not want to disclose such modification information, especially when such information may undermine what the photographer tries

Table 2. An example of modification information contained in a photo's metadata.

Create date	History when	History software	History parameters
2014:07:16 15:13:56	2014:07:19 01:30:03, 2014:08:08 21:17:25	Adobe Photoshop Lightroom 5.4 (Macintosh), Adobe Photoshop Lightroom 5.6 (Macintosh)	Converted from image/x-nikon-nef to image/dng, saved to new location, converted from image/dng to image/jpeg, saved to new location

to convey through the photo. For instance, the contained modification information may cast doubt on the legitimacy of a photo used as digital photographic evidence in court. In addition, celebrities may not like the public to know the photos they were depicted in are actually photoshopped.

2.3 Three Stages of Digital Photos

Based on their propagation process, contemporary digital photos fall into three stages: "fresh," "intact," and "wild." In the "fresh" stage, a photo is freshly taken, free from any post-processing manipulations and still stored in the local camera device. All the metadata information contained in a "fresh" photo is automatically embedded by the camera device, instead of being subsequently introduced by a post processing. In the "intact" stage, a photo has been uploaded online, but remains intact and has not yet been compressed or resized by the hosting media site. For a photo in the "wild" stage, it may have undergone resizing, cropping, and other editing actions conducted by the hosting site, which could change the hidden metadata too. By characterizing digital photos in these three different stages, we aim to depict the status of contemporary digital photos.

3 Fresh Photos

The photos in the "fresh" stage are just freshly created. We examine the metadata information, especially sensitive information, embedded in those freshly taken photos. In this section, we first describe the method used for collecting "fresh" photos and then characterize the collected photos.

3.1 Data Collection

The collection of "fresh" photos is not easy due to their inherent characteristics. We found that it is an effective way to solicit "fresh" photos through crowdsourcing. We posted tasks on a crowdsourcing platform. In each task, the required actions for a worker to take are two-fold: (1) pick up her smartphone, take a photo, and then send the photo to us directly via the instrumented email client application, and (2) take a short survey asking for her demographics information. In addition, to guarantee the unique origin of each photo, each worker is allowed to take our task only once.

Table 3. Demographic statistics of worker participants

Gender	Percent	Country	Percent	Age	Percent	Education	Percent	MobileOS	Percent
Male	71.7 %	India	14.4 %	<=17	2.3 %	Graduate	17.7 %	Android	72.8 %
Female	28.3 %	USA	13.7 %	18–24	45.8 %	Bachelor	47.0 %	iOS	18.2 %
NA	NA	Serbia	7.8 %	25–34	36.3 %	High Sch.	33.3 %	WindowsP	5.2 %
NA	NA	Nepal	5.3 %	35–44	10.8 %	Middle Sch.	1.7 %	Blackberry	1.8 %
NA	NA	Macedonia	4.4 %	>=45	4.7 %	Elementary	0.4 %	Other	2.0 %

For each received photo, we employed various methods to check if it is freshly taken with a smartphone rather than a photo randomly grabbed from the Internet. In addition, according to our tests, sending a photo via email does not affect its embedded metadata. Thus, our task requirements guarantee that the collected photos are freshly created and intact from any post-processing manipulation. The data collection lasts for two months and we collected 782 photos in total. We filtered out 170 photos that are either post-processed or created by other tools. We use the set of the remaining 612 photos for our study. We address potential ethical concerns on our data collection in Appendix A.

3.2 Characterizing "Fresh" Photos

Demographics. The 612 photos were collected from 612 unique workers from 76 countries. Table 3 lists the demographic statistics of the worker participants: (1) 71.7 % of workers were male and the rest were female, (2) 45.5 % of workers were from the top five countries, including India, United States, Serbia, Nepal, and Macedonia, (3) 82.1 % of workers were between the ages of 18–34 and 10.8 % between 35–44, (4) 47 % of workers received the bachelor's degree, 33.3 % with high school degree, and 17.7 % with graduate degree, and (5) 72.8 % of photos were taken with Android phones and 18.2 % with iOS phones.

(Sensitive) Metadata Prevalence. Although Table 1 lists quite a few metadata fields typically embedded in a photo, a specific photo often has a large portion of its metadata information missing. According to our measurement results, we found that two metadata fields, camera make and model, are the most fundamental metadata information. That is, if they are missing in a photo, most other metadata fields are missing too. Thus, we decide whether a photo contains metadata information based on these two fields. A photo is regarded as containing metadata if either of the two fields has a non-empty value.

With the help of a third-party library [2], we examined the prevalence of metadata information among 612 "fresh" photos. We also examined if "fresh" photos contain any sensitive metadata fields, including geolocation, owner's information, and modification history, as mentioned in Sect. 2. Figure 1 shows the percentages of photos containing metadata and sensitive metadata fields. As high as 86.4 % of "fresh" photos contain metadata, which demonstrates the prevalence of metadata information among freshly taken digital photos. As of the sensitive metadata fields, 15 % of fresh photos are tagged with geolocation information. The results show that although nearly all smartphones are now

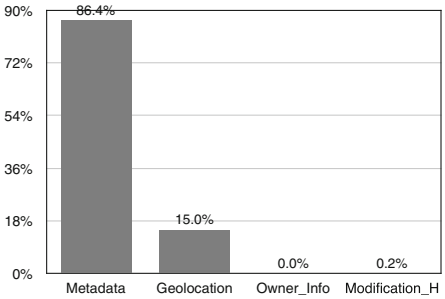

Fig. 1. Percentage of "fresh" photos containing metadata information.

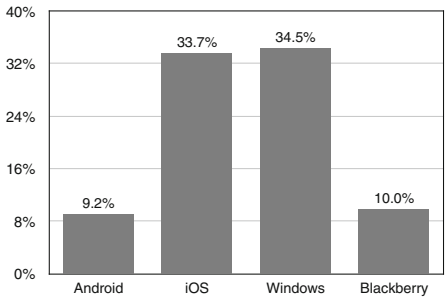

Fig. 2. Percentage of "fresh" photos tagged with GPS for smartphone OS.

GPS-equipped, only some of them are GPS-enabled. The percentage is expected to be even lower if more people are aware that smartphones may automatically embed geolocation into photos and then choose to turn the GPS functionality off. None or hardly any of "fresh" photos contain photographers' information or modification history in their metadata. We speculate that it is due to (1) our strict task requirements and (2) the possibility that these two kinds of sensitive metadata fields may not be automatically embedded at the time of a photo shot.

Impact of Smartphone OS on Geolocation Metadata. It is interesting to examine which kind of smartphone OSes are more likely to automatically embed the sensitive geolocation information into photos. Figure 2 shows that about one third of iOS and Windows phones automatically embed geolocation into photos while only about 10 % of Android and Blackberry phones do this.

4 Intact Photos

In the "intact" stage, photos have been posted online while retaining intact metadata information. From this perspective, "intact photos" could reflect the status of metadata in digital photos at the time of being shared online. In this section, we describe our data collection method for "intact" photos and examine the embedded metadata information in them.

4.1 Data Collection

To collect such photos, we crawled photos from Flickr, a large photo-sharing website, using its API with the download option of "original size," which guarantees that the photos remain original and intact from the site. More specifically, we collected two sets of "intact" photos from Flickr. The first set denoted by *Flickr_p* contains 18,404 photos exclusively taken with smartphones. Those photos were crawled from the Flickr group "Smartphone Photography" where all photos were taken with smartphones. The other set denoted by *Flickr_6* contains 43,704 photos uploaded within six months from July 1, 2014 to December 31,

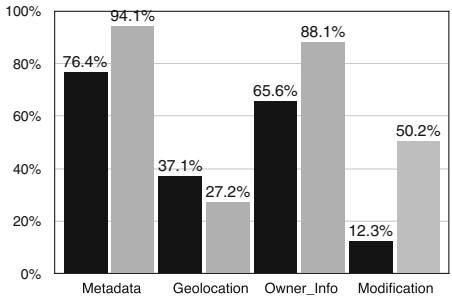

Fig. 3. Percentage of "intact" photos containing metadata information. In each of four pairs of columns, the left black column represents *Flick_p* while the right gray *Flick_6*.

2014. Our further examination shows that 94.3 % of the photos in *Flickr_6* were taken with digital cameras.

4.2 Metadata Information Embedded

Similarly, we examined the percentage of "intact" photos containing metadata information, especially sensitive metadata fields. Figure 3 shows the percentages of "intact" photos containing metadata and sensitive metadata fields.

It shows that intact photos in *Flickr_p* and *Flickr_6* have quite high percentages containing metadata information, 76.4 % and 94.1 %, respectively. The results indicate that most digital photos taken with either digital cameras or smartphones contain metadata when being uploaded online. In addition, 37.1 % *Flickr_p* and 27.2 % *Flickr_6* photos contain GPS information. Considering 15 % of "fresh" photos tagged with geolocation, we speculate that some photo owners may embed GPS information into photos during post processing to better show their photographic works on Flickr. Moreover, up to 65.6 % and 88.1 % *Flickr_p* and *Flickr_6* photos contain the photographer information, which could pose a great risk of identity leakage to photo owners. Additionally, about a half of *Flickr_6* photos contain modification information. Most photos in the set are taken with professional digital cameras and photo owners often show intense interest in refining their works with image processing software. By contrast, a much lower percentage of *Flickr_p* photos taken with smartphones are modified.

5 Wild Photos

In the "wild" stage, most online photos have lingered on the Internet for a while and may have experienced multiple modifications by the hosting sites. In this section, we attempt to figure out the metadata information remaining in the "wild" photos and explore how the top media sites handle the photos hosted on them.

5.1 Data Collection

We employed two methods to collect "wild" photos. The first method is to randomly collect photos by Google Images Search. In the custom search control panel, we set the image type as photo, file type as JPG/JPEG files, image size as larger than 400*300, and the date range from January 1, 2012 until January 1, 2015. Nearly all digital photos are in JPEG format. The specified image size can filter out most of graphs, drawings, and other non-photo images. In addition, we only focus on the photos posted online in the past three years. We totally collected 38,140 photos in this way and denoted them by *GoogleImage*.

Secondly, to investigate top media sites' policies on handling photos, we need to obtain a representative set of media sites. Alexa categorizes millions of sites and defines a list of site categories [4], from which we selected seven categories, which are "social networking," "weblog," "news," "college," "government," "classified," and "shopping". The reason why we chose them is that presumably the sites in these categories usually host large amounts of photos. Alexa provides for each category a list of the top 500 sites. We selected the top 100 sites for each category and thus we had 700 unique top ranked sites in total as our subject representative of online media sites.

Not every photo appearing on a site is hosted by the site. A photo is considered being hosted on a site only if its image URL has the same domain as the site URL. Only the photos hosted on a site are eligible to be used for studying the site's polices. During our photo collection from each site, we only crawled the photos hosted on that site. Specifically, for each of the 700 sites, we attempted to crawl 1,000 photos that appeared online after January 1, 2012. Those photos are expected to reflect the photo policy used by the hosting site under an assumption that the site has not made significant changes to its photo handling policy in the recent years. Due to unexpected factors including network connection failure and access permission denied, we were able to crawl 97,664 photos from 679 unique sites. To ensure the representativeness of these photos, we filtered out the sites from which less than 10 photos were collected. Finally, we had 97,403 photos for 611 unique sites as our dataset for the study, about 160 photos per site on average. This set of photos are denoted as *TopSitesPhoto*.

Figure 4 depicts the number of photos crawled from each site. It shows that about 80 % of sites have over 60 photos crawled, about 35 % of sites have over 120 photos crawled, and about 20 % have over 300 photos crawled. We crawled a maximum number of 1,026 photos for one site[3].

5.2 Metadata Information Embedded

Figure 5 shows the percentages of "wild" photos containing metadata, especially those sensitive metadata fields. It shows that the percentages of "wild" photos containing metadata information in the sets *GoogleImage* and *TopSitesPhoto* are 41.5 % and 40.4 %, respectively, which are much smaller than that of "intact"

[3] We crawled the site twice and collected over 1,000 photos.

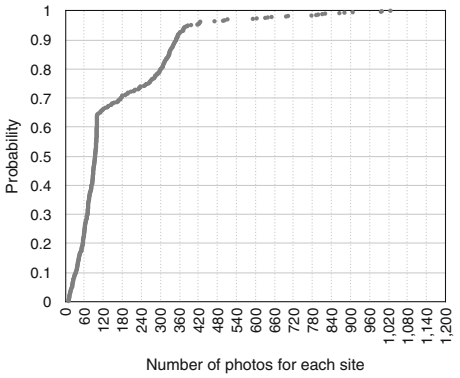

Fig. 4. CDF of number of photos crawled from each site.

photos (up to 94.1 %). In addition, very few "wild" photos are tagged with GPS coordinates. In *GoogleImage* and *TopSitesPhoto*, the percentages are 0.6 % and 1.8 %, respectively, smaller than those of "fresh" and "intact" photos. Moreover, only 13.2 % of *GoogleImage* photos and 8.7 % of *TopSitesPhoto* photos contain photographers' identification information. About 25.4 % of *GoogleImage* photos and 14.1 % of *TopSitesPhoto* photos contain modification history information. These results imply that compared to "fresh" and "intact" photos, a considerable proportion of "wild" photos have their embedded metadata stripped away.

5.3 Inferring Online Sites' Photo Handling Policies

Based on *TopSitesPhoto*, we have built a set of photos for each of the 611 unique sites. We attempt to infer a site's photo handling policy by characterizing the photos collected from the site. Specifically, we aim to answer two questions about a site's photo handling policy. One is whether the site resizes the photos it hosts,

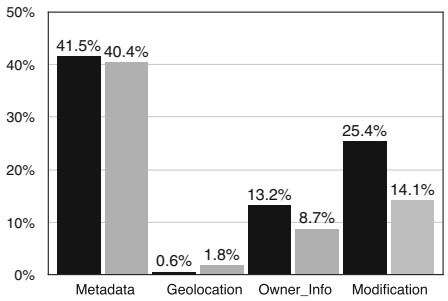

Fig. 5. Percentage of "wild" photos containing metadata information. In each of four pairs of columns, the left black column represents *GoogleImage* while the right gray *TopSitesPhoto*.

and the other is whether the site removes the metadata information embedded in those photos.

Whether a Site Resizes its Hosted Photos? After upload, a photo is typically compressed and resized by the hosting site in several sizes. For instance, Instagram uses an image size of 640 pixels in width and 640 pixels in height for nearly all its hosted photos. More commonly, an online site confines a photo's longest side length to a small set of values. Flickr resizes its photos in the following sizes: 100 pixels (on the longest side), 240 pixels, 800 pixels, 1600 pixels and so on [10]. Therefore, if the majority of photos hosted by a site have their longest side (width or height) lengths falling into a small set of numbers, then we speculate that the site does resize the photos it hosts.

For each photo in our dataset, we retrieved its longest side length from its file information. About 2 % of photos had no image size information available and were ruled out. Suppose "*DDDD*" is the longest side length value that is observed most frequently on a site. We calculated the proportion of the photos on the site with their longest side length of the value "*DDDD*". We then leveraged the proportion number to decide whether the site resizes its photos or not. If over 50 % of photos on the site have the longest side length of "*DDDD*", the site is considered to resize its photos. The argument is based on our observation that among more than 40,000 photos downloaded from Flickr with "original size" option, only 3.47 % have their longest side length of 1,600 pixels, while this length value occurs much more frequently for the photos that have been resized.

Figure 6 shows what percentage of sites that are regarded to resize the photos on their sites across the 7 categories. It is not surprising to see that only 3.0 % of "College" sites and 10.5 % "Government" sites have resized their photos, since colleges and governments usually have sufficient hosting resources to store high-resolution photos. About 36.7 % of "News" sites are estimated to resize the photos they host. A close examination reveals that news sites often resize their photos to many different sizes, which thereby lowers the percentage of photos with a unique longest side length size. In reality, there are probably

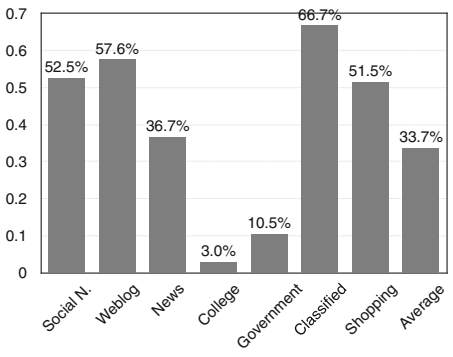

Fig. 6. Percentage of sites estimated to resize their photos across the seven categories.

much more news sites that resize their photos. In each of the other four categories, "Social networking," "Weblog," "Classified," and "Shopping," over 50 % of sites have resized the photos they host. The sites in those categories often contain large amounts of photos and resizing photos is an effective means to save valuable storage space. Irrespective of categories, at least one third of all sites in our dataset are regarded to resize the photos they host. Note that our results represent a lower bound of the percentage of sites that resize their photos.

Whether a Site Strips Out the Metadata Information Embedded in the Photos it Hosts? There is another issue people may be concerned about when they upload photos online. As mentioned before, we use two fields in the metadata—camera make and model—to determine if the metadata information exists or not. For each site in our dataset, we calculated the percentage of its photos containing metadata information. Note that a photo may have its metadata information erased by its owner before posted online. Thus, our estimated percentage of online sites that strip out the metadata information of the photos they host represents an upper bound.

Figure 7 shows the CDF of the percentage of photos containing metadata information on each of the 611 sites in the seven categories. About 16 % of sites have no photos containing metadata information. It is highly probable that those sites remove the metadata information from all hosted photos. About 45 % of total sites have at least half of their hosted photos containing metadata information. We determine that a site adopts a policy of removing photo metadata information if no photos hosted by the site contain metadata information; otherwise, the site is considered to preserve the metadata information of photos it hosts.

Figure 8 shows the percentage of sites in each category which are estimated to preserve the metadata information of photos they host. Again we found that the two categories "College" and "Government" present quite different statistical characteristics in preserving the photo metadata than the rest five categories. Specifically, 98 % of college sites and 93.7 % of government sites are estimated to

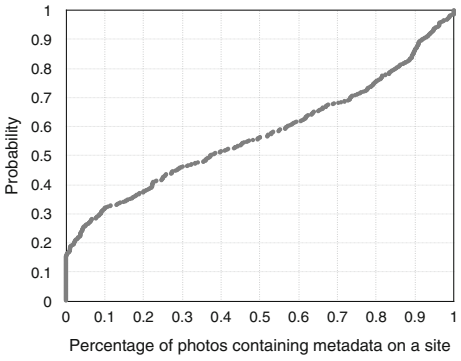

Fig. 7. CDF of the percentage of photos containing metadata information on each site.

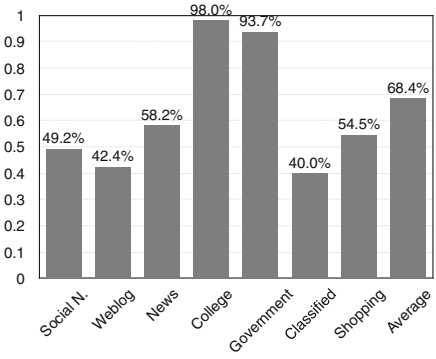

Fig. 8. Percentage of sites estimated to preserve the photo metadata information across the seven categories.

preserve the photo metadata information. Combined with the above estimation results on a site's photo resizing policy, we draw the conclusion that college and government sites seldom resize the photos they host or remove the embedded photo metadata information. In each of the other five categories, the proportions of the sites that preserve the photo metadata information are between 40 % and 60 %, much lower than those of college and government sites. On average, up to 68.4 % of the top sites in the seven categories preserve the photo metadata information, which suggests that a number of online photos may still have their metadata information open to public access for years.

6 Re-identification Attack

Except the sensitive metadata fields including geolocation, owner's information, and modification history, other metadata fields may appear relatively innocent. However, in this section, we demonstrate the feasibility of exploiting a trivial looking metadata field for re-identification attack.

Even without the photographer information explicitly included, a photographer can still be identified based on even only one photo she ever took. This can happen through a new attack vector—the camera serial number field in the photo metadata. A camera serial number can uniquely identify a camera most of the time.[4] All photos taken with a same digital camera are supposed to have the same serial number if provided.[5] In theory, a single photo with a camera serial number embedded could be used to trace other online photos taken with the same camera. Those photos together facilitate identifying the photographer.

We figured out that a public online database *stolencamerafinder* [3] could be leveraged to search for online photos tagged with a given camera serial number, although the online service was established to help find stolen cameras. For each

[4] A serial number is unique within a camera brand. Combined with camera make and model, a serial number can uniquely identify a camera.

[5] Smartphones typically do not store their serial numbers in their photos.

given serial number, *stolencamerafinder* returns a list of online photos taken with the same camera, and for each photo provides the page URL where the photo is posted and the image URL linking to the photo.

Next, we do experiments to prove it quite easy to identify a photo owner with only one photo she ever shared online in the case that the photo has a camera serial number embedded. About 12 % of the "wild" photos in the two sets *GoogleImage* and *TopSitesPhoto* were found to contain the serial number information. We randomly selected 2,000 unique serial numbers from them, then manually searched each serial number in the *stolencamerafinder*, and finally got back search results for 1,037 serial numbers in total. Note that not every camera serial number could get search results back. For those 1,037 serial numbers, by following the image URLs returned, we collected 38,140 photos that were posted on 4,712 unique websites. The photos collected for a specific serial number only represent a subset of all photos available online and tagged with the same serial number, due to the impossibility of finding all online photos with a given serial number.

Figure 9 shows the cumulative distribution function (CDF) of the number of photos that a single serial number links to. About 30 % of serial numbers link to over 25 photos and about 10 % link to over 100 photos. The average number of photos linked to a same serial number is 36.8, the median is 10, and the maximum is 923. With the considerable number of photos tagged with a same camera serial number, together with the page URLs where the photos are posted, and the photos already existing in the photo sets *GoogleImage* and *TopSitesPhoto*, we were able to set up a knowledge base for each serial number (tentatively a digital camera). The rich information available can evidently disclose much more privacy information about the camera owner than a single serial number itself. This demonstrates the potential of a camera serial number as an attractive attack vector for mounting privacy attacks.

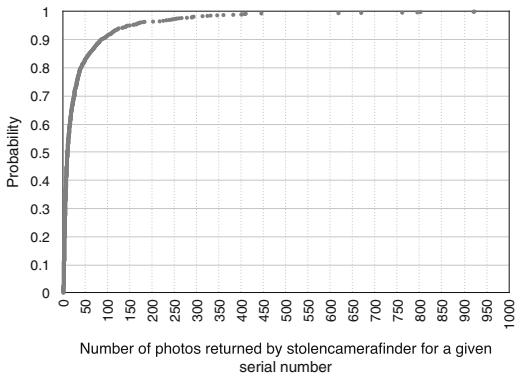

Fig. 9. CDF of the number of photos returned by *stolencamerafinder* for a given serial number.

Table 4. List of the information typically contained in an account profile in each of the five OSNs. Note that the listed information represents the maximum amount of information available with public permissions of an OSN account.

OSN	Account profile information
Flickr	Name, Occupation, Living City, Hometown, Gender, Personal Website(s), Email, Joined Time, Biography, Age, Religion
500px	Name, Biography, Living City, Contact, other OSN accounts
Google+	Name, Gender, Living City, Colleges Attended, Current Employer, Work Experience
Twitter	Name, Occupation, Living City, Telephone, Email, Personal Webpage(s), Joined Time, Photos and Videos, Tweets, Followings, Followers and Favorites
Facebook	Name, Living City, Gender, Education, Telephone, other OSN accounts, Life Events

Identifying a Photographer. The page URL and the page where a photo is posted can provide important clues to reveal a photographer's online identity. For instance, the URL https://plus.google.com/XYZ/photos suggests that the photographer should have a Google+ [8] account with the ID of "XYZ". Following the URL allows us to retrieve more information about the photographer, such as her real name, college attended, current employer, and photos posted on her account page. We have observed a great many such URL strings in our dataset with photographers' online social networks (OSNs) account IDs embedded. The involved OSNs include but not limited to Flickr, Facebook [6], Twitter [7], Google+, and 500px [9]. A photographer may have her multiple OSN accounts disclosed in this way. Table 4 lists the information typically contained in an account profile of the five social networks mentioned above. It shows that an account profile typically contains demographics and other sensitive information including age, gender, education, occupation, living city, other OSN accounts, and much more. Once one OSN account is identified, the true identity of the user in the real world can be readily disclosed.

Figure 10 shows the percentage of serial numbers from which we are able to identify the corresponding camera owners' IDs in one or more OSNs by scrutinizing the page URLs where the photos were posted. Among the 1,037 unique serial numbers in our dataset, 51.4 % (533) of the serial numbers have the camera owners' OSN accounts identified, and 9.0 % (93) have account IDs in two or more OSNs identified. And for one serial number we even identified the camera owner's four account IDs in four OSNs respectively.

As mentioned before, we were able to retrieve about 37 online photos on average for a given serial number. Those photos tagged with the same serial number may contain metadata information that could help identify the photographer. We closely examined the metadata information embedded in the related photos for each of the remaining 504 serial numbers without any OSN accounts

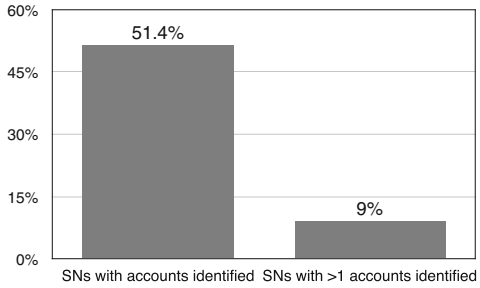

Fig. 10. Percentage of camera serial numbers (SNs) with camera owners' OSN accounts identified.

identified in the previous step. Among them, we successfully identified the photographers for 116 serial numbers. Compared to the photographers with their OSN accounts identified, the available information on those 116 photographers are restricted to the photo metadata embedded, mainly including their names, the processing softwares, and OSes used. However, more information could be collected online once a person's name is identified. Overall, 62.6 % (649) of serial numbers have had their photographers identified.

7 Discussion

One goal of this work is to track the propagation of the sensitive metadata information embedded in the digital photos at different stages. One ideal way is to monitor the process of creation, modification, and elimination of the metadata information contained in a same set of photos that sequentially experience three stages—"fresh," "intact," and "wild." However, it is very hard to obtain such an ideal photo set in large-scale. Instead, we employed different data collection methods and obtained three kinds of photo sets to represent the digital photos at the corresponding three stages.

We collected 612 valid "fresh" photos through crowdsourcing in a period of two months. Each photo collected was taken by a unique participant with a unique device, and participants from 76 countries contributed to this dataset. In addition, those photos were solicited directly from smartphones and no photos taken with digital cameras were collected in order to avoid data contamination. Therefore, although the dataset size of "fresh" photos is not comparable to those of "intact" and "wild" photos, its representativeness is high enough for this study.

To infer online media sites' policies on handling metadata information in the photos they host, we adopt a passive approach, that is, by examining the metadata information of the photos collected from the sites. Actually, we once considered to take an active approach to detect media sites' policies, by submitting (uploading) different types of photos to the sites, then re-downloading them, and comparing metadata fields. However, we had to abandon this approach because most of the 611 sites in the seven categories have specific user groups and are not open to public registration, not to mention photo uploading.

Table 5. Main functions of the browser extension prototype

Sensitive metadata	Potential threats	Website's policy
Geolocation	Location disclosure, house robbery	Metadata removing
Photographer's name	Identity disclosure	Photo resizing
Modification history	Undermining photo's authenticity	NA
Camera serial number	Re-identification attack	NA

Although it is known that a camera serial number can uniquely identify a camera to some extent, we are not aware of any previous research work revealing potential threats arising from this attribute in an empirical and systematic manner. We demonstrated the feasibility of re-identification attack by exploiting camera serial number. We were able to identify over 60 % of photo owners based on their camera serial numbers available in a public online database.

When a user shares a digital photo online, two questions about privacy issues are readily raised. One is whether sensitive hidden metadata information is embedded in the photo. The other concerning question is what the media site will do with the photo. According to our experiment results, a considerable proportion of digital photos contain sensitive metadata information, and many sites resize the photos they host or remove the embedded photo metadata information. In our future work, we will develop a browser extension to give users direct answers to these two questions.

The major functions that the tool should have are illustrated in Table 5. Specifically, once the sensitive metadata information in a photo being uploaded is detected, the browser extension should issue an alarm by popping up a window on the screen and provide customized alert information, including the sensitive metadata information embedded, the corresponding privacy risks, and the current visiting site's policy on photo handling. Note that the browser extension should display the alert information only when the privacy-related metadata information is detected, and thus it should not often interfere with normal photo upload workflows. Although there are already browser extensions for photo metadata visualization, we will focus on informing users of the sensitive metadata contained and customized privacy risks. Moreover, we will ensure users' right to know the actions that the hosting media sites will perform on their photos.

8 Related Work

Several previous works conduct user studies to understand users' privacy decisions during the photo sharing process and their privacy concerns on others' photo-sharing activities. Clark et al. [11] revealed the problem of unintended photo storage without users' awareness, which is mainly caused by the automatic features of cloud-based photo backup services. Ahern et al. [12] found that mobile users' decisions to post photos privately or publicly were determined more by identity or impression concerns than security concerns. Besmer et al. [13] made

similar findings. They studied users' perception of being tagged in undesired photos uploaded by others. They found that a user's privacy concerns on that domain were mainly related to identity and impression management within her existing social circles. Henne et al. [14] showed in their survey results that among the information potentially disclosed by the tagged photos, personal references and location data raised most privacy concerns.

More related to our work, several researchers examined the privacy threat posed by the textual metadata information contained in online photos. Friedland and Sommer [15] focused on the privacy threats posed by the geolocation information available online. They showed that the geolocation data could be exploited to mount privacy attacks using three scenarios on Craigslist, Twitter, and YouTube, respectively. Pesce et al. [17] demonstrated that photo tagging on Facebook could be exploited to enhance prediction of users' information like gender, city, and country. Another work from Mahmood and Desmedt [16] discussed possible privacy violations from Google+'s policy that any users who access a photo can see its metadata online. While the above three works addressed the privacy issues with photos, we investigated the privacy issues with online photos on a much larger scale. We assessed the privacy risks arising from leakage of all possible sensitive metadata information rather than just geolocation data. Moreover, our study is not restricted to one media site. Instead, we collect our photo dataset from hundreds of top-ranked websites and through crowdsourcing platforms. Those photos cover various stages, i.e., "fresh," "intact," and "wild." In addition, we introduce a new attack vector and show its unexpected power in conducting a re-identification attack. We also performed a large-scale measurement of photo handling policies adopted by various categories of media sites.

Another large body of previous work has attempted to enhance people's privacy when sharing photos online. Besmer et al. [22] designed a privacy enhancement tool to improve the photo tagging process on Facebook. The tool allows tagged users to negotiate online with the photo uploaders about the permission settings on the photo. Fang and LeFevre [18] built a machine learning model for OSN users to configure privacy settings automatically with a limited number of rules provided. Zerr et al. [23] developed privacy classification models for users to search for private photos about themselves posted by others at an early stage. Henne et al. [21] proposed a watchdog service that allows users to keep track of potentially harmful photos uploaded by others at the expense of sharing their location data with the service. Ra et al. [19] presented a selective encryption algorithm that enables a photo to hide its "secret" part from the host photo-sharing site and the unauthorized viewers and only expose its "public" part. Ilia et al. [20] refined the access control mechanism currently used by OSNs on photo sharing. The new mechanism allows the depicted users in a photo to decide the exposure of their own face, and could present photos with the restricted faces blurred out to a visitor. Complementary to those works attempting to enhance privacy on the web server side, this study assesses the privacy risks arising from sensitive photo metadata and provides some guidelines for developing client-side privacy leakage prevention tools, which should be able to alert online users of

potential privacy risks posed by uploading photos and also inform them of the photo handling policies adopted by the currently visiting website.

To the best of our knowledge, we have conducted the first large-scale empirical measurement study of the status of contemporary digital photos at the three different stages. In addition to examining the sensitive metadata information embedded, we inferred the photo handling policies used by hundreds of top-ranked sites, and proposed to exploit the camera identification number as an attack vector for re-identification attack. We are not aware of any previous work studying these topics.

9 Conclusion

In this paper, we performed a data-driven assessment of privacy risks on contemporary digital photos. We first collected from the Web nearly 200,000 digital photos at three different stages as our dataset. Then for photos at each stage, we measured the prevalence of metadata and assessed the privacy risks posed by metadata leakage. We found that metadata is quite prevalent among digital photos at each stage. In particular, 15 % of "fresh" photos, about 30 % "intact" photos, and about 1 % "wild" photos were tagged with GPS coordinates. The percentage of "wild" photos containing other sensitive metadata information is also much lower than that of "intact" photos. A possible reason is that online sites often remove the metadata information of the photos they host. Our speculation was confirmed by our investigation of photo handling policies based on nearly 100,000 photos crawled from 679 top sites in seven categories. We further found that photo policies used by a site vary with the category that the site belongs to. Finally, we proposed to use the camera serial number as a new attack vector towards privacy inference and demonstrated its power in deriving both online and real-world identities of a photographer with just one photo she ever took. In our future work, we will build a browser extension prototype to prevent users' photo privacy leakage and increase their knowledge of the online services' policies on photo handling.

Acknowledgement. We would like to thank our shepherd Chris Kanich and the anonymous reviewers for their insightful and detailed comments. This work was partially supported by ARO grant W911NF-15-1-0287 and ONR grant N00014-13-1-0088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

A Ethical Consideration

In our study, we leveraged several methods to collect photos, including: (1) soliciting "fresh" photos from crowdsourcing workers, (2) crawling photos from Flickr using its API, (3) random Google Image Search, and (4) crawling top websites for limited amounts of photos. Note that our crowdsourcing study has been vetted and approved by the Institutional Review Board (IRB) at our institution.

During our photo collection, we did not receive any concerns or get warnings from those involved sites and did not interfere with their normal operations. In addition, with the collected photos, we anonymized the metadata information embedded before using them for study. We strictly abide by the copyright licenses if present.

References

1. Number of photos uploaded to Flickr. https://www.flickr.com/photos/franckmichel/6855169886/
2. ExifTool library. http://www.sno.phy.queensu.ca/~phil/exiftool/
3. Site stolencamerafinder: Find your camera. http://www.stolencamerafinder.com/
4. Alexa top sites by category. http://www.alexa.com/topsites/category/Top
5. McAfee's location is leaked with photo metadata. http://www.wired.co.uk/news/archive/2012-12/04/vice-give-away-mcafee-location
6. Facebook: https://www.facebook.com/
7. Twitter: https://twitter.com/
8. Google+: https://plus.google.com/
9. 500px: https://500px.com/
10. Flickr file size limits. https://www.flickr.com/help/photos/
11. Clark, J.W., Snyder, P., McCoy, D., Kanich, C.: I saw images I didn't even know I had: understanding user perceptions of cloud storage privacy. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI) (2015)
12. Ahern, S., Eckles, D., Good, N., King, S., Naaman, M., Nair, R.: Over-exposed? Privacy patterns and considerations in online and mobile photo sharing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI) (2007)
13. Besmer, A., Lipford, H.R.: Poster: privacy perceptions of photo sharing in facebook. In: Proceedings of the 4th Symposium on Usable Privacy and Security (SOUPS) (2008)
14. Henne, B., Smith, M.: Awareness about photos on the web and how privacy-privacy-tradeoffs could help. In: Adams, A.A., Brenner, M., Smith, M. (eds.) FC 2013. LNCS, vol. 7862, pp. 131–148. Springer, Heidelberg (2013)
15. Friedland, G., Sommer, R.: Cybercasing the joint: on the privacy implications of geo-tagging. In: Proceedings of the 5th USENIX Conference on Hot Topics in Security (HotSec) (2010)
16. Mahmood, S., Desmedt, Y.: Poster: preliminary analysis of Google+'s privacy. In: Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS) (2011)
17. Pesce, J.P., Casas, D.L., Rauber, G., Almeida, V.: Privacy attacks in social media using photo tagging networks: a case study with Facebook. In: Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (PSOSM) (2012)
18. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: Proceedings of the 19th International Conference on World Wide Web (WWW) (2010)
19. Ra, M., Govindan, R., Ortega, A.: P3: toward privacy-preserving photo sharing. In: Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI) (2013)
20. Ilia, P., Polakis, I., Athanasopoulos, E., Maggi, F., Ioannidis, S.: Face/Off: preventing privacy leakage from photos in social networks. In: Proceedings of the 22nd ACM Conference on Computer and Communications Security (CCS) (2015)

21. Henne, B., Szongott, C., Smith, M.: SnapMe if you can: privacy threats of other peoples' geo-tagged media and what we can do about it. In: Proceedings of the 6th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec) (2013)
22. Besmer, A., Lipford, H.R.: Moving beyond untagging: photo privacy in a tagged world. In: Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI) (2010)
23. Zerr, S., Siersdorfer, S., Hare, J., Demidova, E.: Privacy-aware image classification and search. In: Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR) (2012)