Sherlock Holmes' Evil Twin: On The Impact of Global Inference for Online Privacy

Gerald Friedland
ICSI
fractor@icsi.berkeley.edu

Gregor Maier
ICSI
gregor@icir.org

Robin Sommer
ICSI/LBNL
robin@icir.org

Nicholas Weaver
ICSI
nweaver@icsi.berkeley.edu

ABSTRACT

User-supplied content—in the form of photos, videos, and text—is a crucial ingredient to many web sites and services today. However, many users who provide content do not realize that their uploads may be leaking personal information in forms hard to intuitively grasp. Correlation of seemingly innocuous information can create inference chains that tell much more about individuals than they are aware of revealing. We contend that adversaries can systematically exploit such relationships by correlating information from different sources in what we term *global inference attacks*: assembling a comprehensive understanding from individual pieces found at a variety of locations, Sherlock-style. Not only are such attacks already technically viable given the capabilities that today's multimedia content analysis and correlation technologies readily provide, but we also find business models that provide adversaries with powerful incentives for pursuing them.

Categories and Subject Descriptors

K.4.4 [**Computers and Society**]: Electronic Commerce; K.4.2 [**Computers and Society**]: Social Issues

General Terms

Security

Keywords

Online Privacy, Global Interence, Anonymization, Security

1. INTRODUCTION

> *"I have the advantage of knowing your habits, my dear Watson," said [Holmes]. "When your round is a short one you walk, and when it is a long one you use a hansom. As I perceive that your boots, although used, are by no means dirty, I cannot doubt that you are at present busy enough to justify the hansom." "Excellent!" I cried. "Elementary," said he. "It is one of those instances where the reasoner can produce an effect which seems remarkable to his neighbor, because the latter has missed the one little point which is the basis of the deduction.*
>
> —Watson and Holmes in "The Crooked Man"

Correlation of seemingly innocuous information can create inference chains that tell much more about individuals than they are aware of revealing. Consider this example: Public records indicate you own a house. A friend's photo taken at a party you gave, posted to Flickr with exact GPS coordinates, reveals others who also attended. Among them, face recognition software identifies a guest recently arrested for a drug offense [31]. Given these correlations, others might conclude that you associate with known convicts, with unpredictable implications for your reputation. Combining such scenarios with the web's tendency to "never forget," we have entered a world where minor indiscretions—and even innocuous personal facts subject to misinterpretation—can live on forever [11].

No wonder that the new generation of "social web services," such as Facebook, Flickr, and Foursquare, raise the concern of privacy activists. When Google jump-started "Buzz" (their now defunct social network that drew upon Gmail contacts), they "faced a firestorm of criticism" [34]. Facebook's introduction of "Places" likewise drew significant criticism for its default settings, which make it easy to reveal *someone else's* location [55]. In some countries, Google's StreetView has even prompted lawmakers to debate "Google Laws" outlawing comprehensive photography of whole neighborhoods [45].

It can be hard to separate hype from actual risk in such discussions. Still, online privacy research has clearly not yet sufficiently addressed protecting users from unexpected harm. In particular, the discussion so far has all but ignored an area that is poised to open up a whole new class of powerful privacy attacks: automated content analysis enabling cross-site correlation of personal information.

User-supplied content—in the form of photos, videos, and text—is a crucial ingredient to many web sites and services today. Many are adding "social features" for rating and sharing at a fast pace, often with direct interfaces to the major social networks such as Facebook and Twitter.

However, many users providing content do not realize that their uploads may be leaking personal information in forms hard to intuitively grasp. Publishing a variety of heterogeneous information across sites and services creates a cloud of information that, taken together, can reveal a quite comprehensive picture of a person, even if individually none of the pieces may be worrisome on their own. As a simple example, having photos of the same person on separate profile pages immediately links the involved accounts. Likewise, hearing the same speaker in different videos may lead from an innocuous professional setting to an embarrassing recording.

We contend that adversaries can systematically exploit such relationships by correlating information from different sources in what we term *global inference attacks*: assembling a comprehensive un-

NSPW'11, September 12–15, 2011, Marin County, CA, USA.

derstanding from individual pieces found at a variety of locations, Sherlock-style. Not only are such attacks already technically viable given the capabilities that today's content analysis and correlation technologies readily provide, but we also find business models that provide adversaries with powerful incentives for pursuing them.

Unfortunately, the privacy controls that web services offer today are quite ineffective against such attacks. They typically operate at the level of coarse sharing permissions, granting the right to see information to some users but not to others. However, as they do not consider the actual *content* of what is being shared, these simple policies provide no help with limiting the unanticipated impact of global inference. Consider the objective of keeping one's current location private: when in Paris, one probably knows better than publishing a picture of the Eiffel tower on Facebook. However, what about the photo having no visual clue of Paris but showing somebody *else* who has just announced his location on Twitter? We argue that individuals are extremely unlikely to take such considerations into account, while at the same time there are no technical protections in place that help them assessing their impact.

To develop support for users, we need to better understand inference capabilities and their impact on privacy. The primary conceptual challenge here concerns assessing the *trade-off* between the benefits that providing personal information to web services offers, versus the risks that doing so entails. While users value a personalized experience—for sites such as Facebook, that is the very reason for using them in the first place—they also tend to underestimate the implications for their privacy. Making matters worse, commercial web sites often have incentives to collect personal information comprehensively, both for individualizing their services and for attracting income via targeted ads [50]. We deem it crucial to analyze both benefits and risks from a user's perspective by assessing an individual's capabilities to understand, define, and enforce their personal "privacy sweet-spot."

The remainder of our discussion digs further into motivation and examples for global inference attacks, including evidence we collected on their feasibility. If not stated otherwise, the basic threat model we assume is that of an external adversary accessing publicly available information for learning specifics about victims who are unaware of providing them, either directly or indirectly. In §2 we start by discussing a fictional yet realistic example of exploiting global inference in a legally gray area, and then discuss related work in §3. In §4, we then look more systematically at attackers capabilities and potential "business models". §5 examines specifically the potential of automated content analysis systems for finding and linking seemingly unrelated information, and §6 discusses privacy protection in the face of such attacks. We conclude in §7.

2. SHERLOCK GOES CRAWLING: AN EXAMPLE

We start our discussion with an example scenario. Consider the following business: Fred works for Schooner Holdings and wishes to gain (possibly illicit) inside information on future profits at the chipmaker Letin. Fred hires Eve, who runs an "expert network". Eve puts Fred in touch with Bob, a Letin employee. In the process of consulting for Fred, Bob is encouraged to reveal information about Letin's upcoming products.[1]

Currently, Eve's greatest limit is *finding* experts like Bob who (perhaps unknowingly) possess potential insider information and are willing to act as consultants. Eve would greatly improve her business if she could find "corruptibles": individuals in the business of interest who might be favorable to legitimate or illegitimate offers.

Thus Eve starts searching social networks in search of individuals who are compatible with her desired level of (il)legality. She instructs her crawler to begin with LinkedIn and web searches, crawling the names and contact information for personnel at companies of interest. Then her crawler shifts to Facebook, Twitter, other social networks, and blogs, beginning with all candidates found in the first pass. This crawler does not just look at the candidates but also at friends of candidates.

She also searches any media, including images and videos, for links to other people that the social network might not provide directly. Face recognition for example can provide probable connections to other profiles. She also examines media for any compromising material, such as illegal acts, drug paraphernalia, or party photos with Charlie Sheen. Eve knows that her automated content analysis does not need to be perfect: she leverages crowdsourcing services like Mechanical Turk [4] to validate potential candidate matches using human labor at a very low cost.

Eve's crawler also queries further public and semi-public records. There are commercial services that map an email address to a mailing address, such as `spokeo.com` or `pipl.com`. Her crawler uses these to discover where candidates live and how much their property is worth (e.g., by using Zillow's access to property tax data and sales history).

With all this data, Eve's crawler can now create "inference chains" which estimate the probability that any given candidate in her set has a potential weakness, enabling Eve to search for possible points of corruption. An individual who is dating someone with a reputation as a gold digger or who purchased their house at the height of the real estate bubble, might have financial problems. Such candidates could be honestly corrupted by offering consulting positions, allowing Eve to expand her expert network within the bounds of the law.

Eve might also contract with those operating outside the law. Then blackmail becomes an attractive option, especially if considering guilt by association. Someone with a security clearance may be vulnerable if his associates are drug abusers, or if he is having an affair that can be inferred through social patterns.

The inference chains produced by Eve's crawler are probably flawed with many false matches. This however does not pose a problem for her, as she can easily verify the crawler's inferences manually. If it takes Eve a minute to validate an inference chain, Eve will find three to four candidate matches a day even if just 1 % of the results are true positives. With each corruptible worth a substantial sum to Eve's business, her labor is highly cost effective.

Eve can also apply further, and purely legal, inference analysis to a corporation as an entity. The US Army already recognizes the effect of subtle information leakage, down to the timing and quantity of food orders [51]. How much valuable information is leaked about corporate activities when the whole of the company is examined: synchronized lack of vacation, comments by significant others about travel plans, the presence of individuals from different companies at corporate parties, and other seeming trivia?

Nothing in the preceding scenario is unrealistic: every step Eve takes can be constructed using today's technology. It is simply a matter of putting all the pieces together to collect and analyze the reams of data which exist on today's social networks and other databases. Unfortunately, there is also hardly any protection against somebody like Eve in place.

[1]This practice is possibly illegal but exists in a gray area and is seemingly routine practice. The Galleon insider trading trial [10] was based largely on the use of expert network consultants.

3. SHERLOCK'S FRIENDS: RELATED WORK

Even though one already finds web sites offering cross-site correlation of user information as a commercial or government activity, the full capabilities of global inference have not yet seen much attention in either the academic community or by private enterprises. However, many past efforts have examined the potential of extracting personal information out of individual data sets using more narrow local inference techniques. Many of these approaches are quite relevant for our discussion in the sense that the individual techniques deployed for either attacking privacy or defending against threats, might be well suited as individual *links* in global inference chains. We summarize some of this work in the following.

Correlation across different data sources is already used as a marketing tool. For example, IBM's *Business Analytics and Optimization* offers social network analytics for banking and financial industries [23], thereby extending traditional data-mining approaches to human-readable data with the aim of identifying opinion leaders and social connections from different channels. Furthermore, global inference chains have received attention as a way of "connecting the dots" [49] for identifying terrorist threats.

Ad networks like DoubleClick track users using HTTP cookies to obtain user profiles for better ad targeting. In addition, many web sites rely on third-parties like Google Analytics to provide them with detailed usage statistics for their sites; which in turn enables these providers to track the surfing behavior of individual users. Such analytics sites have become quite significant: more than 1 % of all HTTP requests originating at our small research institute are due to the master Google Analytics script reporting back to Google what pages people are visiting. Facebook's "Like" button offers Facebook a similar capability as it records not just who "Likes" a page but who *visits* a page.

While the scientific community has investigated correlation between different data sets in terms of privacy implications, most of these efforts have focused on de-anonymizing or compromising a single data set with the help of auxiliary information. In 1997, Sweeney [47] showed that anonymously published medical records can be de-anonymized when correlated with external data, triggering a large body of follow-up work on designing anonymous statistical databases as well as understanding their limitations [2, 13–15, 48]. Narayanan et al. present an algorithm and proof for de-anonymizing sparse datasets [32]. They apply their algorithm to anonymized Netflix movie ratings: given knowledge of a subset a person has rated (e.g., learned from a lunch conversation or public ratings), the system is able to identify *all* movies in the database that the user has rated. In [33], the same idea is used to de-anonymize a social network graph by leveraging a graph from a second network with real identities as auxiliary data. Researchers from Parc investigated inference using web search engines in order to analyze whether anonymized (or obfuscated) private documents that are going to be released publicly can be de-anonymized [9,46]. They do not consider multimedia content nor inference between information that is already publicly available. Nevertheless, their approach can be a valuable tool for building global inference chains. Griffith et al. [20] correlate public birth, death and marriage records from the state of Texas to derive the mother's maiden name of more than 4 million Texans. Balduzzi et al. [5] automatically query 8 social networks with a list of 10 million e-mail addresses to retrieve the associated user profiles. They then correlate that profile information across the networks and are able to identify mismatches between them. More generally, Bishop et al. [7] discuss the need to go beyond "closed worlds" when sanitizing a data set and consider external knowledge explicitly.

Another area of related research is locational privacy. The Electronic Frontier Foundation published an overview of locational privacy aspects [8]. Locational privacy in vehicular systems, e.g., toll collection, is addressed in [22, 41]. Zhong et al. [58] present protocols for secure privacy preserving location sharing. The upcoming HTML 5 standard will include APIs to query a client's location. The *Cree.py* [21] application uses geolocation data from social networks and media hosting services to track a person's movements.

In a recent effort [17], we analyzed the privacy implications of *geotagging*, i.e., high-accuracy location information attached as meta-data to audio, image, and video files. Specifically, we examined the risk that such geotags pose for what we termed "cybercasing": using online data and services to mount real-world attacks. The work was based on the observation that an extensive and rapidly growing set of online services is already collecting, analyzing, and integrating geoinformation. We presented three scenarios demonstrating the ease of correlating geotagged data with publicly available information compromising a victim's privacy. First, we examined tracking a specific person, in this case TV show host Adam Savage. Images posted to his twitter feed allowed us to pinpoint the location of his home and studio. Then we used Craigslist to find *For Sale* classifieds containing geotagged images but no address in the textual description. A fair amount of the geotagged postings offered high-valued goods, such as diamonds apparently photographed at home; making them potential targets for burglars. Finally, we demonstrated how one can semi-automatically identify the home addresses of people who normally live in a certain area but are currently on vacation using video they have posted on YouTube.

Several recent studies have examined privacy protections for web users. Shankar et al. [44] and Yue et al. [56] present an automatic HTTP cookie management system to find an optimal per-domain policy for accepting / denying cookies. Aggrawal et al. [3] analyze "private browsing" modes in modern web browsers and find that *(i)* the definition of what is supposed to be kept private differs widely between browsers, and *(ii)* all current implementations have flaws that leak information. Krishnamurthy et al. [28–30] analyze privacy loss in web browsing and social networks. For social networks they analyze what profile information leaks, how privacy settings affect the leakage and they report on the default settings used by social networks. Wondracek et al. [54] demonstrate that social network users can be fingerprinted using their group memberships, allowing malicious sites to identify users via "history stealing". Other work [6, 57] addresses further privacy issues in online social networks. Jagatic et al. [25] show that phishing attacks [26] have a significantly higher success rate when the victim's social context is considered.

Several web sites highlight the potential of information leakage users might not be aware of. `Sleeptime.org` estimates sleep patterns of Twitter users. `Stolencamerafinder.co.uk` crawls for digital camera serial numbers in online photos in order to find pictures taken with stolen cameras. `Icanstalku.com` publishes geotags found in tweets, and `pleaserobme.com` used status updates form social networks to locate users who were currently not at home but had published their home address.

As one attempt to standardize privacy protections, the W3C *Platform for Privacy Preferences (P3P)* initiative [40] enables web sites to express their privacy policy in a machine-readable form so that browsers can inform users about a site's policy and automatically apply a user's privacy preferences. However, P3P relies on web sites accurately specifying their policy; there are no enforcement

mechanisms. In addition, the P3P working group has suspended its work in 2007 as it found insufficient support from browser implementations [40].

Finally there are commercial services aiming at protecting users' privacy and identity. Reputation Defender [42] offers subscription services that allows individuals or businesses to track information published about them. Reputation Defender also offers attempting to correct or delete incorrect or embarrassing information; they have successfully done so in a particular high-profile case [35]. Identity Theft 911 [24] focuses on business-to-business identity theft management solutions as well as data breach handling and defense.

4. UNDERSTANDING THE EVIL TWIN

When considering the actual risk of falling victim to global inference attacks, one might wonder whether there are indeed sufficient incentives for adversaries to develop the necessary correlation techniques. In the following, we generalize our example scenario from §2 by first categorizing attackers by their capabilities (§4.1) and then discussing possible business models more systematically (§4.2). Seeing these, we predict that sophisticated inference attacks will indeed become a common method of choice, not unlike in recent years the evolution of malware has been driven to a large degree by a well-organized underground economy.

4.1 Types of Attackers

In order to understand the threat model for inference attacks, we start by constructing a taxonomy of attacker capabilities and objectives, classifying them along the following dimensions:

Resources. Attackers differ by the resources they have at their disposal, such as storage, bandwidth, and computational power. An *individual attacker* wishing to construct a global inference attack is often limited in both storage and computation: they can only fetch and query data from external databases, can only do so at a rate of a few Mbps to perhaps a few hundred Mbps, and only have access to a small number of computers.

An *institutional attacker* has significant resources, including potentially petabytes of storage, many Gbps of available bandwidth, and massive amounts of computation, including possibly systems designed specifically for these classes of problems such as the Cray XMT supercomputer [12].

Finally, there is an in-between point, a *moderate resource attacker*. Such an attacker either has a few thousand dollars to spend on cloud computing services or has access to a botnet. Such attackers can have a large amount of storage, bandwidth, and compute power, but only for a limited period of time.

Database Access. The capabilities of an attacker also depend on the data it has access to. With *full data access*, the attacker has access to the complete set of desired information. This allows to use whatever indexing or data transformation the attacker requires. Such databases may be available for purchase by attackers (e.g., Twitter's "firehose" costs $30,000/month [1, 52]); or might simply be downloaded through crawling, i.e., fetching all entries which may be of potential interest.

Almost as powerful is *well-indexed access*. In this model, the attacker does not have access to the database, but the particular indices and search tools are an effective substitute. For example, there is no full data access to Google's search database, but due to the excellent indexing, an attacker only needs to construct suitable queries.

With *poorly-indexed access*, a database may have the information the attacker desires but is indexed in an unsuitable manner. For example, many counties in the US provide online access to property owner information but index only by property address, not owner, making searching for a target's house difficult. Such databases may still be useful if the attacker is either able to crawl the database (to effectively upgrade to *full data access*) or can use another database as an anchor to use the index.

Finally, there is *private data access*. For example, Google has a huge trove of information, including user search history, browsing history via Google Analytics, and mail history via Gmail. Although such data is ostensibly private, the entity which has acquired it may use it, and third parties might also be able to gain access to the information (via, e.g., business relationships, subpoenas in lawsuits, government warrants, intimidation, or break-ins).

Target Model. An attacker can be targeting a single *targeted individual*. In this case, the individual needs a suitable defense to resist the attack, regardless of the status of the defenses of others. Stalkers and related threats are examples of individual targeting.

An attacker can also be targeting the *easiest K of N* victims. This is the classic parable of the "Bear Race" for an individual possible victim[2], and these cases require less extensive defenses. Often, *easiest K of N* is the easiest to conduct for an attacker, as it can have a large set of potential victims but only needs to succeed in compromising the information concerning a small subset. In our scenario in §2, Eve's search for "consultants" is an example of such a strategy.

One interesting property of both *targeted* and *easiest K of N* attacks is that although *finding* an inference chain might be nearly impossible for a human, manually *validating* the correctness of a given chain is often straightforward. This makes these attacks potentially more powerful, as an error-prone automatic inference procedure can be corrected by human validation of the few candidate inferences. Even for larger candidate sets, an attacker can leverage crowdsourcing services such as Mechanical Turk [4] to validate high-noise components of inference chains.

Finally, there are entities who wish to target *everyone*, including both corporations and governments. Such adversaries do not necessarily intend to exploit the information they find about all targets but may just not know whom to target initially, and then later select subsequent targets meeting the particular objectives.

Interesting instances of the latter category are "Advanced Persistent Threats" [53], and also governments gathering intelligence for law enforcement or counter-terrorism. The latter is often referred to as agencies "connecting the dots", and is arguably the best examined application of building global inference chains today. In that context, "open-source intelligence" refers to the use of "sources openly available to and legally accessible by the public" [43]. However, the specifics of such efforts are often pursued behind closed doors and thus not available to the research community for further assessment. We note that notions of "attackers" and "defenders" depend on perspective here as well as on the trust one has into the entities pursuing the analyses. For our discussion, the "attacker" remains the party developing correlations, even if for benign reasons.

From a research perspective, it seems most fruitful to assume the threat model of an *individual* attacker with access to both *well indexed* and *poorly indexed* data sources, with the specific target model depending on the objectives of a particular approach and the public data available. Doing so allows for developing case studies demonstrating powerful attacks while not requiring a vast amount of resources. It also mimics what most attackers are likely to have access to. Furthermore, since this is the weakest attacker model, if such an attacker can perform an attack, stronger adversaries are

[2]"I don't have to outrun the bear, I just have to outrun *you*."

also able to perform the same attack, although potentially more easily. Where helpful, one might also consider the *moderate resource attacker*, using a modest amount of additional cloud services.

4.2 The Business Case for Attacks

The implications of an individually mounted attack are already substantial. However, the privacy threat of global inference is significantly elevated by a number of actual *underground business models* that can exploit such techniques to a significant effect:

Cyberstalking. There are numerous reasons why someone would want a detailed profile of another, ranging from legitimate (e.g., background investigations prior to hiring) to questionable (screening potential dates) to downright illegal (stalking). Already people use Google in this manner, it would be natural for companies to enhance and automate the process, possibly as a service for others.

Cybercasing. Adversaries can use information from online and multimedia sources to "case a joint" to carry out real world attacks like burglaries. Sites like `Pleaserobme.com` and our previous work [17] explore the potential threats of cybercasing.

Attack Preparation. There are many attacks, such as phishing, targeted malcode, and social engineering, that work best when the attacker has a detailed profile of the victim (see, e.g., [25]). By performing inference, such attackers can thus increase the effectiveness of their attacks through better targeting.

Economic Profiling. There are significant economic advantages to understanding what activity a corporation is planning, be it from a competitor or from someone hoping to make money in the stock market. It is natural to consider how global inference targeting individuals could be used to create a picture of a company or institution's activity. This can be the "single leak" case of first searching for all members of the target institution and then looking more specifically for those who inadvertently leaked important information; and the "correlated leak" case where inference is attempted among the individuals identified with a target institution in order to develop a picture of what the institution is planning.

Espionage Targeting. When targeting espionage against a company or governmental institution, the adversary needs to know *who to target*: what individuals may have potentially exploitable weaknesses, such as money troubles, a gambling habit, vices, or political views which could be exploited; as the objective is often not to compromise the institution directly, but to compromise one or more individuals. Global correlation attacks are naturally very well suited to targeting such activity. The adversary would first construct a large list of possible targets, and then profile each target for potential weaknesses.[3]

Cyberframing. If companies and institutions end up performing "cybervetting" in an attempt to defend against having employees targeted, this would naturally not only cause problems with false positives but would also enable *cyberframing*: an attacker creates malicious additional information poisoning a global inference chain, e.g., by adding in photos on an unrelated photostream that somehow implicate the victim.

5. SHERLOCK WATCHES VIDEO

We now examine further to which degree today's technology for automated content analysis can be leveraged for building global inference chains.

[3]This is an *easiest K of N* attack. For example, a student project at MIT created such a tool to infer sexual orientation from Facebook profiles [27], which would have been particularly well suited to blackmailing career members of the US military if "Don't Ask, Don't Tell" remained in effect.

In the multimedia community, *Multimedia Information Retrieval (MIR)*—i.e., the task of matching and comparing content across databases—has rapidly emerged as a field with highly useful applications in many different domains. Serious efforts in this area can be traced back to the early 1990s when devices like digital cameras and camera phones, combined with progress in compression technology and availability of Internet connectivity, started to change peoples' lives. This rapid technological progress created a strong demand for organizing and accessing multimedia data automatically. Consequently, researchers from different areas of computer science, including computer vision, speech processing, natural language processing, Semantic Web, and databases, invested significant effort into the development of convenient and efficient retrieval mechanisms that target different types of audio and video data from large, and potentially remote, databases.

5.1 Multimedia Information Retrieval

In order to understand potential inference attacks that multimedia retrieval enables, it is crucial to examine the structure of its underlying analysis approaches. These are usually classified into detection, verification, and recognition (sometimes called identification) algorithms. *Detection* algorithms search multimedia files for a certain event or object, returning success if found. Many tasks can be reduced to a detection setting; e.g., face localization, which determines pixels in an image that are a part of a face. A *verification* task matches a given object or event with something learned a priori (e.g., for authentication purposes, a fingerprint maybe compared to a training set of authorized individuals). The output of a verification algorithm is typically in the range between 0 and 1 and correlates with the similarity of the verified entity. Finally, *recognition* systems compare an open set of events or objects with a training set. For example, speaker identification systems match candidate speakers to learned profiles to see who of them they recognize. In practice, many applications can be reduced to recognition tasks, which in turn might often be reduced to verification tasks. For example, image location estimation is commonly implemented by measuring image similarity within a spatially arranged database of photographs.

Detection, verification, and recognition algorithms are applied to acoustic, visual, and textual data. Textual or structured (computer-readable) information accompanying multimedia data, called *metadata*, often dramatically increases the effectiveness of multimedia content analysis.

Many typical visual detection tasks can potentially enable inference attacks, including face detection, person detection, and movement detection. Likewise, acoustic detection research has produced algorithms that can reliably detect speech and music, and even generic acoustic event detection is often accurate enough for retrieval use. Visual verification and recognition tasks that might be of interest are face recognition and generic image retrieval. Relevant acoustic recognition tasks include speaker identification and keyword spotting (based on speech recognition). Multimodal recognition tasks, such as audio/visual speaker/face recognition or multimodal location estimation are also promising candidates.

Consider this simple example for exploiting the capabilities of multimedia technology: An old Facebook friend publishes a photo of a person and tags its face accordingly. The face may then be matched with one frame of an otherwise anonymized introduction video on a specialized dating site, linking the persona and revealing not only the identity of the person but also its circle of friends.

An example that requires more inference is the following: The university of a famous professor has posted about 20 lecture recordings of her on iTunes University. Since she wants keep her family

life private, she never uses her real name on social networking sites and blurs faces on anything she posts for friends and family. Being a security specialist, after all, she knows all the settings and makes sure her Facebook friends are only close family. However, at one point, her teenage nephew posts a video on YouTube showing excerpts of her moving speech at a family wedding ceremony. With the many hours of iTunes university footage, speaker verification reveals her nephew's YouTube user name, which in turn is linked to a Facebook account revealing his real name. Using the nephew's tagged MySpace photos – that he had long forgotten – all the members of the family can be identified, even without some of them having social network accounts. The names link back to geotagged blog posts revealing home addresses and income status of the entire family. Further published videos and Twitter messages of the collected relatives allow to create a detailed profile of the professors' activities, habits, and social connections. Global inference at work.

5.2 Dealing with Retrieval Errors

As discussed above, multimedia content analysis has rapidly progressed in recent years, and its accuracy can be quite high today for specific applications. Some methods have already reached a precision suitable for integration into everyday products, e.g., camera-based smiling-face detection. Computerized language identification is reported to be better than humans [37], and face verification can be tuned to err only in 29 out of 10,000 trials [39].

In general, however, retrieval accuracy is still a problem. Especially matching across random noisy data sources remains challenging, and working with "video in the wild" (as, e.g., found on web sites such as YouTube and Flickr that do not have specific quality control in place) has just recently started to emerge as an important research area [18].

However, as our scenario in §2 demonstrates, even a small percentage of "hits" can lead to a significant number of privacy compromises just as a result of the sheer amount of data available. Similarly, Mechanical Turk or black-market CAPTCHA solvers can perform a validation if the problem is easily expressed as a non-specialist question (e.g., "Is there a bong in this image?" or "Are these two the same person?").

That in turn suggests that even with relatively high error rates, multimedia content analysis techniques can be used effectively for such attacks by using "lop-sided" tuning, for example by favoring low false alarm rates over high hit rates when scanning for potential victims to attack. When used in combination with other inference schemes and human verification, remaining retrieval errors will be quickly weeded out.

5.3 Case Studies

Let's examine two specific multimedia algorithms to better understand how they can support inference attacks, including how appropriate tuning is performed in practice. The first example demonstrates that one can reliably derive location information even without having convenient geotags available; and the second uses automated speaker identification to match individuals across videos.

5.3.1 Estimating Locations Without Geotags

The cybercasing scenarios discussed in [17] demonstrate the threat potential of location information, in particular when published accidentally. However, all scenarios discussed there still rely on the presence of geotags, which we found available in only a few percent of all images and videos, seemingly limiting the damage that can be done. However, to demonstrate the power of global inference, we repeated the study's YouTube scenario that searches for videos recorded by the same user at locations far apart within a short time interval (indicating that the person may be on travel right now, and their home thus unwatched). This time, however, we did *not* rely on geotags but instead used a system for *multimodal location estimation* [18].

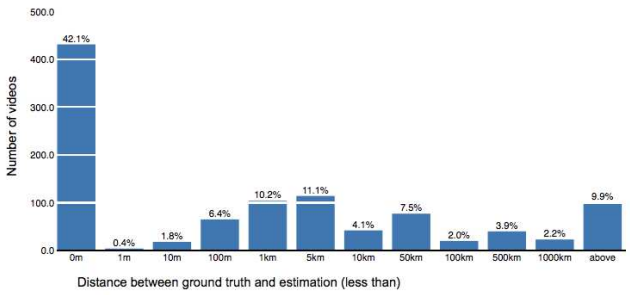

Figure 1: Inferring the origin of a video based on its textual description. In the *MediaEval* placing task, we presented an engine that can determine the location of random Flickr videos based on correlation with other online sources. As the graph shows, over 42 % of the videos were matched with maximum accuracy.

We developed the estimation system originally for an independent project aimed at identifying locations for consumer-produced videos, based on their audiovisual content and textual descriptions. This specific challenge is evaluated annually by the *European MediaEval Placing Task* benchmark [38], in which participants aim to automatically derive latitude and longitude for a random video based on one or more of: textual descriptions (e.g., tags, titles), visual content, audio content, and social information. In addition, the use of further open resources, such as gazetteers or geotagged articles on Wikipedia, is explicitly encouraged. The goal is to come as close as possible to geotags provided by users directly, which are used as ground-truth. The evaluation data set contains 10,000 geotagged Flickr videos with textual descriptions, and 3.2 million geotagged Flickr images.

Figure 1 shows our results on the official dataset. Our system estimated the origin of 1,000 randomly selected Flickr-videos including textual descriptions. It located more than 42 % of the videos with exact accuracy, and 72 % within line of sight (5km) of a typical video camera. The high percentage of exact locations can be explained by the rich detail in textual descriptions. Even if a textual description does not reveal the location per se, it can often be matched with other information from the Internet or the training set that contain geotags or other more explicit location descriptions. For example, a video with the title "My kids Emma and Noah playing in the garden" does not give away any global location information directly if the video is not geotagged. However, scanning a number of different web sites, one finds only a limited set of videos, photos, tweets, blog entries, etc. that contain the words "Emma" and "Noah" and/or "garden" and/or "kids". Our algorithm examines such hits for geotags and other location-explicit textual context. For the latter, we match text against the public online service `geonames.org`, which returns exact latitude and longitude coordinates for over 7 million place names, and further uses Semantic Web technologies to infer locations from sites such as Wikipedia if a description is not found in its core database. Our algorithm then clusters the results spatially, defining the location closest to most others as the desired result. As Figure 1 shows, this rather simple algorithm already yields very precise results.

To repeat the YouTube cybercasing scenario with our location es-

timation system, we used the same initial search keywords as in our original experiment ("kids", "yard", "Berkeley") to find homes in a certain area. We then matched their textual descriptions against a set of two million YouTube video descriptions to find further videos the same user had recently uploaded more than 1,000 km away. Not surprisingly, we were again able to identify a number of cases for potential burglary, this time without the need for any geotags but implicitly determining a user's location from the videos' context. This preliminary experiment already demonstrates the potential of global inference. A more systematic approach for understanding inference possibilities would in addition crawl social networks such as Twitter and Facebook, public databases such as property records, and other openly available data sources. By correlating all this public information, much can be found out about persons without them having published many specifics explicitly.

5.3.2 *Identifying Speakers*

In our second example we discuss tuning a speaker identification system for matching speakers across videos found on general web sites, which is a setting quite different from the more well-defined environments that such systems are normally deployed in. We first briefly summarize background on how speaker identification typically operates, and then discuss the setting relevant for us in more detail. We note that this example is representative of many other content analysis schemes, which will require similar tuning [16].

Most current speaker identification systems have been implemented for the speaker recognition paradigm established by NIST evaluations. These involve a database of *target speakers* and a set of *test utterances*. Each test utterance is matched against each speaker, obtaining a score specifying the likelihood that the sample comes from that speaker. Each such score represents a *trial*. Trials where the speaker matches the utterance are known as *target speaker trials*; trials where it does not, are called *impostor trials*. A threshold across the scores is set, such that scores above the threshold are classified as matches, and scores below as non-matches. Impostor trials wrongly classified as matches are *false alarms*, while target speaker trials classified incorrectly as non-matches are *misses*.

A speaker recognition system that we developed in earlier work participated in the most recent NIST speaker recognition evaluation (SRE10, [36]). In the *clean telephone-telephone* evaluation condition—roughly 2.5 minutes of audio per speaker using clean conversational telephone speech—the system's false alarm rate was only about 0.15 % when tuned for a miss rate of 40 %, meaning that roughly 60 % of all target speaker trials were correctly classified, while only about 0.15 % of the impostor trials were false alarms. Considering a more difficult SRE10 condition—different speakers recorded over different microphones in an interview setting but still with a 2.5 minute sample—with a 40 % miss rate, the false alarm rate became about 0.35 % percent, i.e., still quite low.

Working with audio downloaded from Web sites (such as the audio track of a YouTube video) adds potentially high levels of environmental background noise (e.g., music, traffic, noises from poor microphone quality and audio sampling) to which current speaker recognition systems are less accustomed. However, it is reasonable to assume that many such videos will contain at least small *pockets* of speech that have sufficiently low noise levels. If we consider primarily these for the recognition task (extracted using a speech/non-speech detector), the setting might resemble the NIST *10 sec–10 sec* evaluation condition — the hardest condition evaluated by NIST, i.e., 10 seconds of audio taken from each speaker, using clean telephone channel recordings. For this, our system was able to achieve roughly 5 % false alarms at a 40 % miss rate; and a less than 1 % false alarm rate at a 60 % miss rate.

Returning to our inference application, we are interested in identifying pairs of online videos that share the same speaker. If for each pair, we take (part of) one video's audio channel as the target speaker model and the other one as a test utterance, we can apply the results sketched above. Considering the generally increased noise level, we speculate, based on the noise decrease observed in the 2.5-minute scenario, that at 60 % miss rate the false alarm might be roughly 2-3 %, and at the 80 % miss rate perhaps about 0.3–0.4 %. In other words, using conservative tuning it seems conceivable that for a given YouTube video, we will be able to find 20 % of all other videos having the same speaker, while suffering barely from any false alarms. If we assume that only, say, half of all videos will have sufficiently low noise levels to be processed, we can still identify 10 %.

We note that this a back-of-the-envelope calculation that also depends on further technical issues that we skip discussing here. These include computational performance of the recognition system (which is much faster than realtime for each test-utterance and highly parallelizable), and the need for segmenting the audio channel into pieces having only one individual speaker each (for which there is technology available [19]).

We also emphasize that we do not expect that speaker recognition is used in isolation, but *combined* with other approaches for linking identities, such as analysis of textual descriptions and face recognition. By tuning each scheme for a low false positive rate, we can effectively combine their respective strengths.

6. YOU KNOW MY METHODS, WATSON

We contend that the security and privacy community has not yet paid sufficient attention to threats posed by correlating personal information across site boundaries. Specifically, our domain lacks an understanding of the elevated risks that deployment of state-of-the-art content analysis technology incurs. However, given that a number of business models provide adversaries with incentives to develop such attacks (see §4.2), we deem it crucial to better help users with putting up defenses.

Clearly, we cannot expect to anticipate and counter all possible inference attacks. Even when limiting ourselves to a specific threat, it remains unlikely that we can identify *all* its potential variations. Consider location information: while stripping geotags from photos prevents locating them directly, increasing the sophistication of the attack can overcome that defense, per our discussion in §5.3.1. We argue however that despite this fundamental problem, there is significant practical benefit in helping users to *understand* the impact of their actions, and providing them with tools for *mitigating the risks* they face where possible. Compare this to phishing attacks: Once users understand the risk of revealing their password to a web site that an email asks them to visit, they are in a position to choose not to do so. Along the same lines, once one knows that geotags are attached to a photo, one can remove them before uploading.

Unfortunately, however, when providing personal information online, there is no clear-cut line between good and bad. Furthermore, a user's notion of privacy hardly remains constant over time: While having party photos posted online may be socially acceptable while going to college, they can quickly turn into an embarrassment many years later when looking for a job.

The challenge for the research community is to support users in finding their personal privacy "sweet spots" between the two extremes of not providing anything and just publishing everything. In particular, we believe that bringing potential privacy issues—and even just surprising inferences—to a user's attention is already of great value.

As one concrete example for such an approach, consider today's identify theft services that monitor customer activity, such as credit card transactions, for signs of fraud. In the online world, such a service could take on a new role by *proactively* examining a customer's online sphere for possible correlations leaking personal information in unanticipated ways. Such a system could be operated as a subscription service that continuously watches relevant web sites, building inference chains like a potential attacker would. If the service finds leaks that violate a user's policy, it would notify her, along with further instructions on impact and possible mitigation steps. A simple online version of such monitoring already exists specifically for geotagged photos posted on Twitter: the site *icanstalkyou.com* alerts authors of tweets that come with location information. An extended system could generalize this approach by monitoring for advanced inference potential.

7. CONCLUSION

The security and privacy community has an impressive track record of revealing examples of individual web services and applications leaking sensitive information to the public. However, the next, more fundamental step that is still unexplored, concerns understanding *global inference chains*: what can adversaries derive about individuals by correlating seemingly innocuous public information across independent sources? While simple technical fixes can often stop direct leaks, such chains pose a much larger threat given the *fundamental* difficulty of recognizing the latent possibility of linking seemingly unconnected facts. Global inference has to date received little attention, neither in the public discussion nor in the academic literature.

Specifically, we call attention to recent technological advances developed by the multimedia community, which enable correlation of images, videos, and textual information without requiring the availability of convenient machine-readable meta-data. When combined with modern capabilities to efficiently collect, mine, and correlate large volumes of online information from a variety of sources, such technology opens up the potential for privacy attacks with much more power than today's users realize.

We call out as the first action item for our community to *inform* users about potential risks they face. Enabling individuals to understand what can indeed happen places them in a much better position to adjust their behavior where they deem necessary. Educating the broader public in concrete terms about what is technically possible will help them discover what their already-disclosed information reveals about them. As a concrete step forward, we envision following the spirit of traditional identify theft protection: a novel online service could continuously monitor web resources for information derivable abouts its subscribers, alerting them to potentially sensitive correlation chains as it discovers them.

8. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their thoughtful comments. This work was supported by NSF Award CNS-1065240 and a fellowship within the postdoctoral program of the German Academic Exchange Service (DAAD). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the DAAD.

9. REFERENCES

[1] Gnip. http://www.gnip.com.

[2] AGGARWAL, C. On k-anonymity and the curse of dimensionality. In *Proceedings of the International Conference on Very Large Data Bases* (2005).

[3] AGGRAWAL, G., BURSZTEIN, E., JACKSON, C., AND BONEH, D. An analysis of private browsing modes in modern browsers. In *Proceedings of the USENIX Security Symposium* (2010).

[4] Amazon.com Mechanical Turk, https://www.mturk.com/mturk/welcome.

[5] BALDUZZI, M., PLATZER, C., HOLZ, T., KIRDA, E., BALZAROTTI, D., AND KRUEGEL, C. Abusing social networks for automated user profiling. In *RAID'2010, 13th International Symposium on Recent Advances in Intrusion Detection* (09 2010).

[6] BILGE, L., STRUFE, T., BALZAROTTI, D., AND KIRDA, E. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th International Conference on World Wide Web (WWW)* (2009).

[7] BISHOP, M., CUMMINS, J., PEISERT, S., SINGH, A., BHUMIRATANA, B., AGARWAL, D., FRINCKE, D., AND HOGARTH, M. Relationships and Data Sanitization: A Study in Scarlet. In *Proc. Workshop on New Security Paradigms* (2010).

[8] BLUMBERG, A., AND ECKERSLEY, P. On locational privacy, and how to avoid losing it forever. *Electronic Frontier Foundation*. http://www.eff.org/wp/locational-privacy.

[9] CHOW, R., GOLLE, P., AND STADDON, J. Detecting privacy leaks using corpus-based association rules. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008).

[10] COHAN, P. Why executives risk their job to tip a hedge fund. http://meet-the-street.blogspot.com/2009/10/expert-networks-what-every-iro-needs-to.html, 2009.

[11] COLBERT, S. The Word - Control-Self-Delete, http://www.colbertnation.com/the-colbert-report-videos/351570/august-24-2010/the-word---control-self-delete.

[12] CRAY INC. Cray XMT. http://www.cray.com/products/xmt/.

[13] DINUR, I., AND NISSIM, K. Revealing information while preserving privacy. In *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)* (2003).

[14] DWORK, C. Differential privacy. In *33rd International Colloquium on Automata, Languages, and Programming (ICALP)* (2006).

[15] DWORK, C. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, vol. 4978 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2008, pp. 1–19.

[16] FRIEDLAND, G. Analytics for Experts. *ACM SIGMM Reecords 1*, 1 (April 2009).

[17] FRIEDLAND, G., AND SOMMER, R. Cybercasing the Joint: On the Privacy Implications of Geo-Tagging. In *Proc. USENIX Workshop on Hot Topics in Security* (August 2010).

[18] FRIEDLAND, G., VINYALS, O., AND DARRELL, T. Multimodal Location Estimation. In *Proc. of ACM Multimedia* (October 2010).

[19] FRIEDLAND, G., VINYALS, O., HUANG, Y., AND MÜLLER, C. Prosodic and Other Long-Term Features for Speaker Diarization. *Transactions on Audio, Speech, and Language Processing 17*, 5 (2009), 985–993.

[20] GRIFFITH, V., AND JAKOBSSON, M. Messin' with texas deriving mother's maiden names using public records. In *Proceedings of the International Conference on Applied Cryptography and Network Security (ACNS)* (2005).

[21] H-SECURITY. Cree.py application knows where you've been. http://www.h-online.com/security/news/item/Cree-py-application-knows-where-you-ve-been-1217981.html.

[22] HOH, B., GRUTESER, M., HERRING, R., BAN, J., WORK, D., HERRERA, J.-C., BAYEN, A. M., ANNAVARAM, M., AND JACOBSON, Q. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *MobiSys '08: Proceeding of the 6th International Conference on Mobile Systems, Applications, and Services* (2008).

[23] IBM. Capabilities: Social network analytics for banking and finance. https://www-950.ibm.com/blogs/gbs_business-analytics/entry/capabilities_social_network_analytics_for_banking_finance1?lang=en_us.

[24] IDENTITY THEFT 911. http://www.identitytheft911.com/.

[25] JAGATIC, T., JOHNSON, N., JAKOBSSON, M., AND MENCZER, F. Social phishing. *Communications of the ACM 50*, 10 (2007).

[26] JAKOBSSON, M., AND MYERS, S. *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. Wiley-Interscience, 2006.

[27] JOHNSON, C. Project Gaydar, Sep 2009.

[28] KRISHNAMURTHY, B., MALANDRINO, D., AND WILLS, C. Measuring privacy loss and the impact of privacy protection in web browsing. In *Proceedings of the 3rd Symposium on Usable Privacy and Security* (2007).

[29] KRISHNAMURTHY, B., AND WILLS, C. Generating a privacy footprint on the Internet. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement (IMC)* (2006).

[30] KRISHNAMURTHY, B., AND WILLS, C. Characterizing privacy in online social networks. In *Proceedings of the 1st Workshop on Online Social Networks (WOSN)* (2008).

[31] MCCARTHY, C. N.J. Town posts DUI photos on Facebook, http://news.cnet.com/8301-13577_3-20013632-36.html.

[32] NARAYANAN, A., AND SHMATIKOV, V. Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy* (2008).

[33] NARAYANAN, A., AND SHMATIKOV, V. De-anonymizing social networks. In *Proceedings of the IEEE Symposium on Security and Privacy* (2009).

[34] NEW YORK TIMES. Critics Say Google Invades Privacy With New Service. Feb 13, 2010.

[35] NEWSWEEK. A tragedy that won't fade away, Apr 2009. http://www.newsweek.com/2009/04/24/a-tragedy-that-won-t-fade-away.html.

[36] NIST. 2010 NIST Speaker Recognition Evaluation, http://www.itl.nist.gov/iad/mig//tests/sre/2010/index.html.

[37] NIST. The 2007 NIST Language Recognition Evaluation Results. http://www.itl.nist.gov/iad/mig/tests/lre/2007/lre07_eval_results_vFINAL/index.html, 2007.

[38] PETAMEDIA. 2010 MediaEval Placing Task, http://www.multimediaeval.org/placing/placing.html.

[39] PHILLIPS, P. J. Improving Face Recognition Technology. *Computer 44*, 3 (March 2011), 84–86.

[40] PLATFORM FOR PRIVACY PREFERENCES (P3P) INITIATIVE. http://www.w3.org/P3P/.

[41] POPA, R. A., BALAKRISHNAN, H., AND BLUMBERG, A. VPriv: Protecting Privacy in Location-Based Vehicular Services. In *Proceedings of the USENIX Security Symposium* (2009).

[42] REPUTATION DEFENDER. http://reputationdefender.com/.

[43] SCHAURER, F., AND STÃ¶RGER, J. OSINT Report 3/2010 - The Evolution of Open Source Intelligence. Tech. rep., International Relations and Security Network, ETH Zurich, 2010.

[44] SHANKAR, U., AND KARLOF, C. Doppelganger: Better browser privacy without the bother. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)* (2006).

[45] SPIEGEL ONLINE. Setting Hurdles for Google and Others: Germans Like to Look, But Not to Be Looked At. http://www.spiegel.de/international/germany/0,1518,712485,00.html, 2010.

[46] STADDON, J., GOLLE, P., AND ZIMNY, B. Web-based inference detection. In *Proceedings of 16th USENIX Security Symposium* (2007).

[47] SWEENEY, L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, and Ethics 25*, 2–3 (1997).

[48] SWEENEY, L. k-anonymity: A model for protecting privacy. *Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10*, 5 (2002).

[49] TAIPALE, K. A. Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data. *SSRN eLibrary*.

[50] THE WALL STREET JOURNAL. The Web's New Gold Mine: Your Secrets. Jul 30, 2010.

[51] Army regulation 530-1: Operations security (opsec), April 2007.

[52] VENTUREBEAT. Gnip grabs $2M as it teams up with Twitter in new data selling deal. http://venturebeat.com/2010/11/18/gnip-funding-twitter-data.

[53] WIKIPEDIA. Advanced Persistent Thread (APT). http://en.wikipedia.org/wiki/Advanced_Persistent_Threat, 2011.

[54] WONDRACEK, G., HOLZ, T., KIRDA, E., AND KRUEGEL, C. A practical attack to de-anonymize social network users. In *Proceedings of the IEEE Symposium on Security and Privacy* (2010).

[55] WORLEY, B. Tips to Turn Off Geo-Tagging on Your Cell Phone, http://abcnews.go.com/GMA/video/online-photos-give-privacy-11443667.

[56] YUE, C., XIE, M., AND WANG, H. An automatic HTTP cookie management system. *Computer Networks* (2010).

[57] ZHELEVA, E., AND GETOOR, L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web* (2009), pp. 531–540.

[58] ZHONG, G., GOLDBERG, I., AND HENGARTNER, U. Louis, lester and pierre: Three protocols for location privacy. In *Proceedings of the Privacy Enhancing Technologies Symposium* (2007).