Connecting the Dots: When Personal Information Becomes Personally Identifying on the Internet

Dave Yates, Mark Shute, and **Dana Rotman**
College of Information Studies
University of Maryland
4105 Hornbake Building, South
College Park, MD 20742
{dyates, mshute, drotman} @umd.edu

Abstract

With online social media such as weblogs (blogs), authors seemingly control how much self-identifying information they disclose. However we find that that even authors who wish to remain anonymous will share expressive and access enabling information which, when combined, can be used to positively identify the person. In a case study of three anonymous blogs we demonstrate how to combine investigative analysis with statistical techniques to identify anonymous authors with a high degree of accuracy. Paradoxically, anonymous authors feel as if they can be honest and open with their thoughts and opinions, and thus may be more likely to share more information than they might if their identities were known.

Introduction

When sharing ideas and opinions on the internet, it is almost inevitable that some degree of personal information (information that describes unique characteristics of an individual) is going to be disclosed if for no other reason than to provide context for a comment. A person might disclose their age when a particular movie was released to explain their opinion of that movie. They might reveal the location of their hometown as a reflection of their support for a particular sports team. They might mention their income level to validate a particular political viewpoint. People who seek to disseminate their opinions online may disclose fragments of personal information like this without a second thought to their privacy. After all, such personal information isn't necessarily personally identifying. Or is it?

Social media such as weblogs (blogs) and social networking websites make such sharing of personal information ubiquitous, and as a result many in research and practice are concerned with maintaining the privacy of this information. A primary concern is that personally identifying information (information that positively identifies an individual) such as social security numbers or addresses will inadvertently leak or be disclosed publically; however a growing fear is that 'personal information' (e.g., preferences, recent purchases, family connections) that is not necessarily 'personally identifying information' could, if publically disclosed, compromise individual privacy. In this research, we take the position that personal information can and does compromise privacy; that authors readily disclose personal information even when they consciously do not disclose personally identifying information; and that they are often unaware that this disclosure might lead to others discovering their identities.

To investigate inadvertent exploitation of personal information we conducted a case study of personal information disclosure through blogs as a social medium. Most blogs are published as public websites, unlike typical social networking websites which offer multilevel privacy rules for specific groups or contacts - thus authors control who has access to different information, including personally identifying information. With blogs, authors control access to their personally identifying information simply by not disclosing it in the first place. According to Qian and Scott (2007), roughly 40% of bloggers censor their writing, including by anonymizing their identity. However, in this study we attempt to show that even anonymous bloggers may be identified based on personal information they disclosed, when coupled with other public data sources. After reviewing relevant literature concerning online information disclosure, we present an analysis of three anonymous blogs in which, with the bloggers permission, we attempted to ascertain the authors' identities. Without having to access the authors specific identities, we used investigative procedures combined with statistical analysis to calculate the probability of identifying the individual author from all other people in

the world, based on the information gathered, which for all three cases was greater than 90%. This technique gives researchers and authors insight into the impact on their 'anonymity' from disclosing various bits of information in their blog entries. We conclude with a discussion of the implications of personal information sharing for personal privacy.

Literature Review

Personally identifying information (also called personal data, as specified in the European Union Data Directive 95/46/EC) is defined as personal information that can point to one unique individual. A social security number or a driver's license number is a good example of a single piece of information that is personally identifying, but not all personal information is personally identifying on its own. A person's name is often considered to be personally identifying, but a common name such a "Michael Smith" isn't unique unless it is combined with some other information. Even an uncommon name may be shared by dozens of people around the world. A residential address is often considered to be personally identifying. It may be, if a person is the sole occupant of that residence; if there are multiple residents, the address alone is insufficient for unique identification. The same is true for landline phone numbers. Cell phone numbers however are likely to be unique to one person. Almost any personal information can be personally identifying if it is combined with enough other personal information.

Disclosure of Personal Information

In online environments, the ability of users to maintain their privacy through anonymity is heightened when they control information disclosure (unlike concerns of corporate mismanagement of electronic data). According to DeCew (1997), individual authors manage three types of personal information: self-identifying (such as name, social security number), access enabling (such as address and zip code), and expressive (such as personal interests, experiences, and life situation). Government regulation of personally identifying information is almost exclusively focused on the self-identifying information. Expressive information is, according to Goldie (2006) the foundation of our social relationships and social persona; that is, how we reveal expressive information determines with whom we build social ties online, and how we manage what others think of us. Thus expressive information is commonly shared through online social media such as blogs.

The third (middle) category of information that may be shared is access enabling information. This personal information occupies a privacy grey area, since an address alone does not identify someone, and further individuals routinely disclose access enabling information for purposes of deliveries, or locating resources. Madden et al. (2007) note that the concern of this type of information isn't necessarily it's disclosure as much as its contribution to a lasting and accessible 'digital footprint,' which means the information can be, at some later time, correlated with other information to become personally identifying.

Thus, with blogs we might expect that expressive information is routinely shared, access enabling information may be shared with reservations although the immediate privacy concerns are not necessarily worrisome, and self-identifying information is not shared at all by those authors desiring anonymity. In the next section, we review author options for anonymity and how it affects what information they do share.

Author Anonymity

Online users usually choose among three identification modes – real name, anonymity or pseudonym (Chen et al. 2008). With personal blogs (i.e. blogs written by an individual in the first person as a description of their thoughts and deeds, as opposed to blogs written to convey professional information or in an organizational setting), the latter two forms of identification are much more prevalent than the use of real names. They allow users to conceal segments of their identity, and display any of their "multiple windows" at will (Turkle 1997). Anonymity and self-disclosure are not dichotomous, although some scholars maintain that notion (e.g., Tannen 1998). They are privacy choices that users may use along a continuum of exposure and self disclosure. The ability of users to continuously decide which self-revealing information they wish to disclose allows them to calibrate not only the level of their exposure to others, but also the temporal and situational circumstance in which they will be exposed. It is a complex balance of exposure and disclosure in which the user expresses himself, and provides others with cues as to his identity.

Nissenbaum (1999) asserted, however, that complete anonymity is rare. Authors may be identified through the combination of various properties of themselves, and placed within a smaller set of individuals, ultimately leading to their recognition. As such, anonymity may be viewed as a subjective feeling of untraceability. The more prevalent option is the use of a pseudonym - an arbitrary identifier (e.g. screen-name, user ID) chosen by the user, which may or may not be based on the user's personally identifying information. The use of pseudonym enables users to create an alternative identity that is related to a distinct online persona ("nym") (Froomkin 2003), and reputation, faciliatating continuous interaction with others, under the guise of intentional and partial disclosure of personal information. Pseudonymity allows users to have an identity that is not directly related to their off-line persona, but is rather a form of self-authentication created through some aspect of their identity (location, ID, repeated pattern of actions) (Marx 1999).

Research suggests that anonymous or pseudonymous interaction benefits shy and insecure users (Sheeks and Birchmeier 2007), and allows users to express themselves

more openly and honestly (Qian and Scott 2007), stimulating freedom of thought, self expression, and critical thinking, which are not mitigated by fear of ridicule, social sanctioning or political constraints (Kling et al. 1999). Nissenbaum (1999) construed that the value of anonymity or pseudonymity is in allowing users to participate in social interaction while remaining unreachable, outside the scope of reprisal. Therefore, anonymous and pseudonymous authors may actually share more expressive or access enabling personal information online (Tidwell and Walther 2002), which paradoxically might be used by others to more readily identify the contributor.

Identifying Anonymous Authors

So how much expressive and access enabling personal information does it take to become self-identifying? It depends on the type of information, and the context it is in. A residential address combined with even a first name becomes a unique identifier in most cases. The name of a student's school, combined with their class schedule might be enough to single out an individual. The name of a person's church and the name of their employer might also serve to uniquely indentify a person if you compared the rosters of both organizations. Research has found that people can be personally identified 87% of the time by just their five-digit zip code, gender, and date of birth—all pieces of information generally considered to be non-identifying individually (Samarati and Sweeney 1998).

Samarati and Sweeney's (1998) research was conducted using 135,000 medical records of Massachusetts state employees released to commercial industry and researchers by the Group Insurance Commission (GIC). GIC made a good faith effort to make the data anonymous by stripping the patients' names, social security numbers, addresses and phone numbers out of the records before they were released. The remaining data fields included the patient's zip code, complete date of birth, gender, and ethnicity—along with medical histories—so that medical trends influenced by geography, age, gender and race could be analyzed. At the time, the data was considered to be anonymous.

Samarati and Sweeney purchased a list of Massachusetts state voting records from the state. These public records also included zip codes, birth dates and genders of registered voters, along with their names and addresses. When the two lists were compared, Samarati and Sweeney found that the zip code/birth date/gender trio of personal information disclosed in the released medical records allowed a match to unique individuals on the voter registration list 87% of the time.

This research has been well circulated in information science circles, being cited in hundreds of scholarly works. Sweeney went on to develop a set of industry policies and best practices for releasing data sets that will ensure what she calls the "k-anonymity" of the people being reported on (Sweeney 2002).

In summary, it is well established that individuals can be identified by a sufficient volume of access enabling personal information when it is compiled and released as a structured data set. Can the same rule apply when personal information is self-disclosed in a more informal manner? Can a person accidentally sacrifice their privacy by disclosing personal information online that they thought was non-identifying? We conducted the following study to investigate these questions.

A Research Study of Three Blogs

To determine if we could identify blog authors we conducted a case study in which we manually reviewed three blogs kept by anonymous authors to determine if they disclosed sufficient personal information. Blogs were used because they consolidate information in a single repository that is easy to search. The same study could have been done with microblogs, chat sessions, comment threads, social networking websites, massively multiplayer online role-playing games (MMORPG) chatter or any other internet venue where people may disclose personal information in the process of exchanging ideas in a public forum. These formats would have taken significantly more time to aggregate however, whereas blogs unlike other social media offer an aggregate repository that is easy to review, by researchers and incidental readers alike.

Overview of Blogs Selected for the Study

The blogs chosen for this study were found using Google to search for phrases such as "maintain my privacy" and "remain anonymous" on sites such as Blogspot, Wordpress and Livejournal. Each author was contacted and agreed to participate in the study. All information reprinted from these blogs—author pseudonyms, dates, and quotes—has been altered from its original form for publication in order to prevent quoted searches and assure the author's anonymity. The three blogs involved are:

- *Big Dad's World* by Big Dad, who discusses technology and politics, including his local and church politics, from a conservative viewpoint.
- *The Slut Next Door* by Quirky Slut, who writes primarily about her sexual encounters and the events surrounding them.
- *Elfling's Journal* by Elfling, who keeps friends & family up to date on her life, plans/coordinates activities, and reviews movies, liquor & perfume.

Each of them was chosen because they had declared a desire to maintain their privacy while still disclosing a variety of personal information. These three blogs come from different genres that are representative of a variety of personal blogs in which anonymity might be maintained. Each of the authors write under a pseudonym, a common strategy for maintaining privacy employed by almost a third of the bloggers who responded to one privacy survey

(Qian and Scott 2007). They also use their blogs to share details of their personal life, a characteristic of over half of the bloggers who responded to another survey (McCullagh 2008).

The selected bloggers participated in an interview in which they shared their reasons for writing anonymously, their views about the disclosure of personal information and their understanding of online privacy. Interviews were conducted via email, using the bloggers' publicly shared email addresses. Each blogger then gave us permission to search their blog for access enabling and expressive information. No access enabling information was asked for or provided in the interviews.

Each of the bloggers explained their reasons for writing anonymously. Big Dad chooses to use a pseudonym because he does not want "the random drive-by vicious commenters to have a way to pester me in my personal life." Quirky Slut wants to avoid "people seeking me out to try to say they [had sex with] me." Their attempts to maintain their anonymity through the use of a pseudonym are proactive, seeking to prevent an undesirable event. Elfling on the other hand, chooses to use a pseudonym as a reaction to a previous event in her life. "An employer once gave a customer my full name. [The customer] started calling me at home to harass me."

For all three bloggers, the motivation to remain anonymous comes down to a desire to avoid harassment—a very real possibility, in fact. The organization Working to Halt Online Abuse (WHO@) compiles and publishes statistics on incidents of online harassment and cyber-stalking. They receive reports of 50 to 75 cases every week. According to WHO@'s 2008 statistics, 43% of victims had no prior relationship with their harasser, and 25% of the cases included threats of physical violence. Incidences peaked in 2005 and have since been in decline (WHO@ n.d.). Both Big Dad and Quirky Slut began blogging after this peak, when awareness of the possibility of harassment was high.

In addition to keeping their name a secret, each blogger has certain information they are careful never to disclose on their blog. Big Dad never reveals "Where I live, where I work… I never use last names of individuals other than politicians." Quirky Slut withholds "My college. My work. My family." And Elfling conceals "My husband's name… the names of any children I know… where I work."

Voluntary Personal Information Disclosure

None of the three bloggers mentioned any concern about disclosing their birth date or gender. Big Dad did say that he did not reveal where he lives, but the profile page for his blog lists his hometown, so presumably he was referring to a specific street address rather than something as general as a zip code. All three bloggers indicated that they are careful with names—both their own and other people's—which makes sense; names are the most common personal identifier. When combined with any other personal data that associates a person with a limited group of people—a town, an organization, an event—even a common name is likely to identify only one unique person. This is actually consistent with the findings in McCullagh's (2008) survey in which over half of the respondents said that protecting people's personal information was important. McCullagh did not provide a complete list of the kinds of information the respondents said they kept secret, but the comments she did publish did not include any reference to zip code, birth date or gender.

It's interesting that all three of the bloggers mention their workplace as information that they do not disclose, but perhaps not surprising. Recently there has been a rash of anecdotal evidence about employees being disciplined by their employers because of activities on social networking sites (Moses 2009). In addition to risking the employer's displeasure at being associated with what is written in the blog, disclosing an employer would associate the blogger with a very limited group of people, making them more identifiable.

Investigative Procedure

After interviewing each blogger, their websites were reviewed for access enabling information. In particular, clues about their zip code, birth date and gender were sought, and almost always found. In addition, information about marital status and dwelling type would prove to be useful. To be certain that the details in each blog were factual, the bloggers were asked if they would ever consider falsifying personal information in their blog in order to ensure their anonymity. None of the bloggers in this study claim to employ deception as an anonymity strategy. Elfling admitted that she would lie if she thought it was necessary, but could not recall having done so. Both Big Dad and Quirky Slut claimed a moral disinclination towards lying. So the personal information disclosed in their blogs is assumed to be free of any intentional deceit.

To make use of that information, some sort of comparator list is needed. Samarati and Sweeney (1998) paid for state voting records. Other public records such as driver's license data may also serve this purpose. A proprietary database of personal records might also be used, such as the membership database of a national video rental chain, or the student and alumni records of a large university. Such a use might be against policy or unethical, but still possible.

As this study lacked proprietary access, it made use of AlescoLeads, an online tool provided by Alesco Data Group which contains data on over 200 million consumers "Compiled from multiple data sources such as telephone directories, credit files, mail responders, government records and other proprietary sources…" AlescoLeads allows the user to create a list of consumer addresses for direct mail marketing. The user first specifies a zip code or group of zip codes then selects other demographic data, such as gender and a two-year age range. Birthdays are not filtered, but can be included in the final data set along with names and addresses. The tool reports the number of

records that meet the stated criteria, but requires the user to purchase the list to get the actual data. Our authors did not wish to be identified, and we were primarily interested in the impact of combining information on the chances of positively identifying an author. Thus, we did not access the actual AlescoLeads data but rather employed a statistical formula to calculate the chance of uniquely identifying each author. The formula calculates the probability that only one person on the filtered list of AlescoLeads records has the blogger's exact birthday, culled from the information at hand..

The formula to determine that probability is

$$\left(\frac{d-1}{d}\right)^{l-1}$$

where d is the number of days in the target year or years, and l is the total number of people on the list. This simplified formula makes the assumptions that birth dates are equally distributed throughout the year and are independent of all other factors.[1]

This study makes the assumption that all three of the bloggers are in the AlescoLeads database, and that their information is accurate. The analysis and probability calculation for each individual blogger are detailed below.

Big Dad's World. On his profile page, Big Dad states "I'm in my 60's, a grandfather/husband/Christian/country boy…" revealing his gender right away. His profile page also lists his location as "Angela: Montana: United States." A quick Google search shows that the only zip code in Angela, Montana is 59312. It takes a bit more work to assemble a complete birth date for Big Dad. In his November 6, 2007 entry, Big Dad wrote "I'm 63 years old, and loving life!" giving us an age. Almost a year later on October 27, 2008 following a vacation he wrote "After we got back to the lodge last night…my wife and friends threw a bit of a surprise party for me." giving enough information to reveal Big Dad's full birth date. A party is not necessarily held on the actual birthday, but other content in this entry gives the strong impression that Big Dad's birth date is on October 26, 1944.

The same October entry provides a bit more personal information about Big Dad—he is married, or was less than seven months ago at the time of this writing. There is no mention of a divorce or his wife's passing in later blog entries so it is safe to assume that his marital status remains the same.

Finally, in a post on July 14, 2007, Big Dad wrote of his grandson "He was able to drive by himself by that point, in my little pickup, and he was driving in circles around the house and my mother's mobile home." This anecdote indicates that Big Dad lives in a single family home, as opposed to an apartment building or townhouse. So the following criteria can be used to create an AlescoLeads list:

- Zip Code: 59312
- Age: 64-65
- Gender: Male
- Marital Status: Married
- Dwelling Size: Single Family Home

The returned list includes 66 names of people born over a two year time span. Since 1944 was a leap year, for the equation d=731 and l=66.

$$\left(\frac{731-1}{731}\right)^{66-1} = .914$$

Thus there is a 91.4% chance that Big Dad is the only person on the list with the birth date October 26, 1944. It seems quite likely that a person who bought this data list from AlescoLeads could uniquely identify Big Dad.

The Slut Next Door. In an entry dated June 6, 2007, Quirky Slut confirmed the assumption that she is female when she wrote "Being a girl means I can usually get whatever I want just by flirting." In the FAQ page of her blog she wrote "I live in the Albuquerque, NM area." No more specific geographic detail could be found. The greater Albuquerque metropolitan area is comprised of 44 different zip codes, so by living in a big city and being consistently vague, Quirky Slut is actually doing a pretty good job of protecting her anonymity.

On September 19, 2007 Quirky Slut wrote "My birthday is over. So long teenage years." She had a previous post on September 17 in which she made no mention of her birthday, so September 18 is most likely the day. She most likely turned 20 that year, making her birth year 1987. Later, on March 25, 2009 she confirmed the year when she described an upcoming vacation. "We can really enjoy Las Vegas since we're both 21 now," she wrote.

In her entry on August 19, 2009, Quirky Slut disclosed her marital status when she wrote "I'm not married, nor am I attached to anyone." She revealed her dwelling size on April 19, 2007 when she described her living arrangements, "Well, I sort of live with my parents but I live in an apartment above the garage they used to rent to students." This will actually turn out to be the crucial piece of access enabling information that will yield a high probability of uniquely identifying Quirky Slut. The following Criteria were used to create an AlescoLeads list:

- Zip Code: 44 selected for the entire Albuquerque area
- Age: 20-21
- Gender: Female
- Marital Status: Single
- Dwelling Size: Single Family Home

The returned list included just 72 names. Since neither 1986 nor 1987 were leap years, the equation variables are d=730 and l=72.

[1] We thank Dr. Paul Smith of the University of Maryland Mathematics Department for sharing this formula.

$$\left(\frac{730-1}{730}\right)^{72-1} = .907$$

Thus there is a 90.7% chance that Quirky Slut is the only person on the list born on October 24, 1987. If it had not been for the dwelling size criteria, the list would have contained 484 records yielding just a 51.5% probability of uniquely identifying Quirky Slut. Presumably, most single 21-year-old women in Albuquerque do not live in single family homes, and that distinction makes Quirky Slut easier to identify.

Elfling's Journal. Elfling's blog lacks any kind of profile page or FAQ, which makes it a bit more difficult to isolate the most basic information on her. The fact that the blog has been ongoing since 2002 also makes it difficult to be certain that older references to facts such as marital status and dwelling size are up to date. Nonetheless, the standard personal information can be found. On April 11, 2008 Elfling disclosed her gender with the statement "Also my dreaded female check-up is this week. Going to the doctor always makes me anxious." She has made numerous references to the city of Gastonia and the state of North Carolina throughout her blog, making it easy to assume her hometown. The closest single statement confirming this is on July 2, 2006 when she wrote "You can ride or caravan with us (leaving Gastonia, NC around 8:30am)." Gastonia, NC has five zip codes—not as bad as Albuquerque, but still lacking in precision.

A complete birth date can again be obtained from two separate entries. There appears to be awareness among all three bloggers that disclosing a complete birth date might be too revealing, but if that's the case, they also share an ignorance or forgetfulness of what information they had disclosed previously. Elfling made various references to upcoming or past birthdays over the years, but she pins down the exact day on August 5, 2008 when she wrote "Thanks again to everyone who came to my birthday. My natal day is actually tomorrow, so to celebrate..." Almost a year later on July 9, 2009 she disclosed the year when she wrote "There was some talk of my impending 40th birthday" making her complete birth date August 6, 1969.

Elfling's marital status is revealed in a fairly recent post from November 14 2008 when she wrote "…it is also the Hunter's and my First Wedding Anniversary." The Hunter is a pseudonym frequently mentioned in Elfling's blog. He also has a blog that Elfling frequently links to. Confirming that he is her husband offers a second source of access enabling information that can be mined.

The most recent post that discloses a dwelling size is from June 19, 2008 when Elfling wrote "I'm so glad to report that the heating and cooling system is working in our condo again." Later posts will discuss attempts to buy a home, and very recent posts at the time of writing discuss efforts to get approved for a mortgage, but there is no mention of actually completing a purchase or moving in.

Elfling's zip code is still uncertain. While Quirky Slut was more easily identified because few unmarried 21-year-old women in Albuquerque live in single family homes, the same is not likely to be true for 40-year-old married women living in condominiums in the suburbs. Fortunately, The Hunter's blog narrowed down the zip code by describing an incident in which he had to walk home from work on April 23, 2008. "I just had to get home… I followed Catawba Creek through the golf course... I turned right when I got to the tracks and kept walking." Looking at Google maps shows that the train tracks that cross and then run north (a right turn from the municipal golf course) of Catawba Creek form the border of only two zip codes. The following Criteria were used to create an Alesco list:

- Zip Code: 28052 or 28054
- Age: 40-41
- Gender: Female
- Marital Status: Married
- Dwelling Size: Multi Family Home

The returned list included just 26 names. Since neither 1969 nor 1970 were leap years, the equation variables are d=730 and l=26.

$$\left(\frac{730-1}{730}\right)^{26-1} = .966$$

Thus there is a 96.6% chance that Elfling is the only person on the list born on August 6, 1969. If it had not been for the information found on her husband's blog, the list would have contained 65 records yielding a 91.6% probability of uniquely identifying Elfling, about the same as Big Dad. Incidentally, Elfling's husband discloses his first name in his blog, which could be used to further confirm Elfling's identity.

Discussion

While the sample size for this study was not large enough to provide conclusive results, it does demonstrate that personal information casually disclosed online can be used to uniquely identify a person. In each of the three cases studied, the authors could be identified with greater than 90% probability, despite their efforts to limit disclosures and remain anonymous.

At least for the three blogs included in the case study, the results are disturbing for authors who take advantage of anonymity (or pseudonymity, more appropriately) to freely express their thoughts and actions, especially important moments in their lives such as birthdays and events with family members. Although the authors studied were scrupulously careful to control self-identifying information, they were much more open with access enabling and expressive information. However, blogs, as with many other social media, are a ready archive for all of this information. Thus authors must be mindful not only of what personal information they share in each particular post, but the sum total of information shared on the entire blog.

Identifying an anonymous person by sifting and combining information is, right now, a labor intensive effort that requires analytical and associative thinking, so it is less likely that a computer program could be written to identify anonymous authors and invade their privacy on a large scale. However efforts have surfaced recently for search engines to index social networking information from websites such as Twitter (http://twitter.com) and Facebook (http://www.facebook.com) for real-time search and sentiment analysis. The danger is to the anonymous person who is targeted by an identification search for one reason or another. While there may be a legitimate reason to uncover an anonymous identity, it is far easier to imagine this technique being used to stalk or harass someone. However, as more and more sources of data are made available online and organizations like businesses, higher education, and governments maintain increasingly larger and more detailed data stores, the opportunities for automating data correlation among blog entries and separate sources will increase.

Health informatics researchers (c.f. Krishna et al. 2007, Sengupta et al. 2008) are particularly concerned with data combination and patient privacy given recent efforts to make patient records available online. Although laws such as the Health Insurance Portability and Accountability Act (HIPAA) regulate personally identifying information disclosure, this research raises new questions as to the extent of safeguards necessary to prevent patient identification based on a medical condition or treatment regimen.

Efforts to educate people about the significance of the zip code/birth date/gender combination may help them to increase their own privacy. Further research could be done to survey a wider sample of bloggers and other online authors to get a more precise idea of what kinds of information they believe it is safe to disclose or not, McCullagh (2008) provides an excellent analysis of bloggers concerns based on how they want to be perceived by others. We suggest that research should also investigate whether or not authors would share personal stories or opinions if they knew that information might one day be used to positively identify them.

Of course there are other combinations of personal information that can be just as dangerous as the zip code/birth date/gender trio depending on what type of record set is accessible for comparison. In this study, marital status and dwelling size were invaluable in identifying two of the participants because AlescoLeads allowed that criteria to be filtered. A student who mentions the name of their college, might easily be identified by a person with access to a database of students and alumni. A person who mentions picking up a movie at a particular video rental chain might easily be identified by a clerk at one of those stores. Future educational efforts should take the variety and range of personal details that can be personally identifying into account when addressing privacy literacy.

In fact, any kind of personal information can probably be access enabling if the author is targeted by someone with access to the right set of data records. But the majority of all blogs are used as personal journals (Herring et. al. 2006). How personal can a journal be if you can't mention the town you live in, how old you are, your gender, whether you're married, where you went to college, where you shop or any of the other details that might be exploited? Bloggers might consider employing more fine-grained access control to their blogs, such as a post-by-post decision on public vs. restricted readership list.

Walther and colleagues (c.f. Tidwell and Walther 2002, Walther 1996) have explored social media as a 'hyperpersonal' environment, meaning that authors share details about themselves them would be hesitant to do in a face to face setting, because it is necessary when information we typically use to identify others such as appearance and gestures are missing. This phenomenon, when combined with the longevity of information shared and stored in social media, suggests that even those authors who are concerned with protecting their privacy should be careful of what they write and to whom.

New strategies for policing large databases of personal information may be needed. Another topic for further research might be a survey of owners of such databases to determine what, if any, safeguards exist to protect against their misuse. Unfortunately, while organizations that maintain large sets of such data may have some sort of policy in place for its proper use, it will probably take some well publicized misuse of such a database before adequate security becomes widespread.

Conclusion

In the end, the responsibility for anonymity online comes down to the author, who must weigh the amount of personal information they wish to disclose against the perceived threat of being targeted by someone who wishes to identify them. The less personal information they disclose, the more anonymous they are. So long as large databases of personal information like AlescoLeads are available, it would seem that anyone who mentions their hometown, their birthday, their age and their gender can be identified. It's just a matter of connecting the dots.

References

Alesco Data Group. 2009. Retrieved May 2009 from http://www.alescoleads.com/ index1.cfm.

Big Dad. 2009. Various entries. Big dad's world. Retrieved April-May 2009, from [Redacted].

Chen, H-G., Chen, C. C., Lo, L., and Yang, S. 2008. Online privacy control via anonymity and pseudonym: Cross-cultural implications. *Behaviour & Information Technology* 27(3): 229-242.

DeCew, J. 1997. *In Pursuit of Privacy: Law, Ethics & the Rise of Technology*. Ithaca: Cornell University Press.

Elfling. 2009. Various entries. Elfling's journal. Retrieved April-May 2009, from [Redacted].

Froomkin, M. 2003. Anonymity in the Balance. In Nicoll, C., Prins, J.E.J. and Van Dellen, M.G.M. (eds.) *Digital Anonymity: Tensions and Dimensions*. The Hague, Netherlands: TMC Asser Press: 5-45.

Goldie, J. 2006. Virtual Communities and the Social Dimension of Privacy. *University of Ottawa Law & Technology Journal* 3(1): 133-167.

Herring, S. C., Scheidt, L. A., Kouper, I., and Wright, E. 2006. Longitudinal content analysis of weblogs: 2003–2004. In M. Tremayne (Ed.), *Blogging, Citizenship, and the Future of Media*. London: Routledge: 3-20.

Kling, R., Lee, Y.C., Teich, A. and Frankel, M.S. 1999. Assessing Anonymous Communication on the Internet: Policy Deliberations. *The Information Society* 15: 79-90.

Krishna, R., Kelleher, K., and Stahlberg, E. 2007. Patient confidentiality in the research use of clinical medical databases. *American Journal of Public Health* 97(4): 654-658.

Madden, M., Fox, S., Smith, A., and Vitak, J. 2007. Digital Footprints - Online identity management and search in the age of transparency. *Washington DC: Pew Internet & American Life Project*. Retrieved December 2009 from: http://pewresearch.org/pubs/663/digital-footprints.

Marx, G.T. 1999. What's in a Name? Some Reflections on the Sociology of Anonymity. *The Information Society* 15(2): 99-112.

McCullagh, K. 2008. Blogging: self presentation and privacy. *Information & communications technology law* 17(1): 3-23.

Moses, A. 2009. Social not-working: facebook snitches cost jobs. Retrieved April 30, 2009 from http://www.theage.com.au/news/technology/web/social-notworking-facebook-snitches-cost-jobs/2009/04/08/1238869963400.html.

Nissenbaum, H. 1999. The Meaning of Anonymity in an Information Age. *The Information Society* 15: 141-144.

Qian, H. and Scott, C.R. 2007. Anonymity and Self-Disclosure on Weblogs. *Journal of Computer-Mediated Communication* 12: 1428-1451.

Quirky Slut. 2009. Various entries. The slut next door. Retrieved April-May 2009, from [Redacted].

Samarati, P. and Sweeney, L. 1998. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. *Technical report SRI-CSL-98-04*. SRI computer science laboratory. Palo Alto, CA.

Sengupta, S., Calman, N.S., and Hripcsak, G. 2008. A model for expanded public health reporting in the context of HIPAA. *Journal of the American Medical Informatics Association* 15(5): 569-574.

Sheeks, M.S. and Birchmeier, Z.P. 2007. Shyness, Sociability, and the Use of Computer-Mediated Communication in Relationship Development. *CyberPsychology & Behavior* 10(1): 64-70.

Sweeney, L. 2002. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 557-570.

Tannen, D. 1998. *The argument culture: Stopping America's war of words*. New York: Ballantine Books.

Tidwell, L. C., and Walther, J. B. 2002. Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: getting to know one another a bit at a time. *Human Communication Research* 28(3): 317-348.

TGWH. (2009) Various entries. Oliver's journal. Retrieved May 2009, from [Redacted].

Turkle, S. 1997. Multiple Subjectivity and Virtual Community at the End of the Freudian Century. *Sociological Inquiry* 67(1): 72-84.

Walther, J. B. 1996. Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Communication Research* 23: 3-43.

WHO@ - Working to Halt Online Abuse. (n.d.) Online Harassment/Cyberstalking Statistics. Retrieved May 2009, from http://www.haltabuse.org/resources/stats/.