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. For example, a person might disclose their age when a particular movie was released to explain their opinion of that movie. 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, is it?

Social media such as weblogs (blogs) and social networking websites make 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 although 40% of blog authors hide their identity (Qian and Scott 2007), many readily disclose personal information even when they consciously do not disclose personally identifying information; and that they are often unaware that this disclosure might let others uncover their identities.

Literature Review

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 is almost exclusively focused on self-identifying enabling information. However, 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.

Online users usually choose among three identification modes – real name, anonymity or pseudonym (Chen et al. 2008) – although with personal blogs the latter two forms of identification are much more prevalent. Nissenbaum (1999) asserts, 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. 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. Research suggests that anonymous or pseudonymous interaction allows users to express themselves more openly and honestly (Qian and Scott 2007). Therefore, anonymous and pseudonymous authors may actually share more expressive or access-enabling personal information online which paradoxically might be used by others to more readily identify the contributor.

So how much expressive and access enabling personal information does it take to become self-identifying? 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 purchased a list of Massachusetts state voting records from the state and compared them to anonymized Group Insurance Commission records. 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.

A Research Study of Three Blogs

To determine if we could identify blog authors based on self-disclosed information we conducted a case study in which we manually reviewed anonymous blogs (twelve were contacted and three agreed to participate) to determine if they disclosed sufficient personal information. The three blogs (and their pseudonymous authors) 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.

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." 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. However, none of the three bloggers mentioned any concern about disclosing their birth date or gender.

Investigative Procedure

After interviewing each blogger about their anonymity choices and writing motivations, and obtaining their permission, their blogs 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. To make use of that information, some sort of comparator list is needed. Samarati and Sweeney (1998) paid for state voting records. 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. However as this study lacked proprietary access, we used AlescoLeads, an online tool which contains data on over 200 million consumers compiled from various sources. 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 names but rather employed a statistical formula to calculate the chance of uniquely identifying each author from a filtered list of consumers, based on demographic criteria (e.g., age, zip code). The formula calculates the probability that only one person on the filtered list of AlescoLeads records has the blogger's exact birthday.

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, and that our bloggers are in the AlescoLeads database.[1]

Big Dad's World. 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

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

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.

Thus there is a 91.4% chance that Big Dad is the only person on the list with the birth date October 26, 1944.

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

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.

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

Thus there is a 90.7% chance that Quirky Slut is the only person on the list born on October 24, 1987.

Elfling's Journal. 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 many as Albuquerque, but still uncertain.

A complete birth date can again be obtained from two separate entries. 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 – offering a second source of access enabling information.

Elfling's zip code is still uncertain. 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.

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 in regards to 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, or 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 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 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, recently search engines have begun indexing social networking information from websites such as Twitter (http://twitter.com) and Facebook (http://www.facebook.com) for real-time search and sentiment analysis. Conceivably, an anonymous person might be targeted by an identification search

Efforts to educate people about the significance of the access enabling information, and specifically zip code/birth date/gender combination, may help them to increase their own privacy. Further research might 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. 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 (McCullagh 2008).

The majority of 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 shop or any of the other details that might be exploited? Social media is considered 'hyperpersonal' (Tidwell and Walther 2002) when authors share details about themselves they would be hesitant to divulge in a face to face setting. Often it is necessary to do so to convey communicative cues when 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. 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.

References

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

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.

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.

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

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.

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.

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.