

# Privacy issues with blogging

**Konstantin Tchernov**

University of Auckland

38 Princes Street, Auckland, New Zealand

ktch003@ec.auckland.ac.nz

+64 27 468 1842

## ABSTRACT

Blogging is an easy way for people to publish logs about anything they want on the Web. While there are many types of blogs, one highly popular use is to keep blogs about personal lives. For social bloggers revealing their private information and thoughts to the world could mean that the boundary of separation between their public and private lives could be broken. This could incur a wide variety of problems ranging from social disasters or employment issues to legal liabilities, or perhaps even physical danger for the younger bloggers. While anyone can control what information they put on their blog, many people still end up publishing on impulse without too much thought of what the consequences may be. Bloggers should be able to strike a suitable balance between sharing their thoughts and breaking their privacy. The growing utilisation of blogs continues to create an increasing need for efficient privacy protection – as the access to this information becomes easier and more widespread. This review will discuss the privacy issues associated with personal blogging, as well as possible solutions that have been attempted to help bloggers maintain their desired balance of disclosure.

### Author Keywords

Internet, web, blogging, social networking, privacy.

### ACM Classification Keywords

H.5.3. [Information Interfaces and Presentation]: Group and Organization Interfaces: web-based interaction, organizational design. K.4.1 [Computer and Society]: Public Policy Issues: Privacy.

## INTRODUCTION

Blogs, originally known as web-logs, are a new communications medium that has emerged over the past decade. Typically, blogs provide a free and easy way for anyone to publish their own content on the web, which is then automatically displayed in a chronological order. No web programming knowledge is necessary and user interfaces are usually simple to use. There are many different types of blogs in many different genres and on many different topics. Some people may want to use blogs as public diaries in order to share their thoughts, keep up with friends, or to even make new friends on the Web; others might have impersonal blogs featuring reviews of products, or discussing their opinions on

what's going on around the world, some bloggers even do their own reporting and present independent news. As opposed to web forums, blogs are designed to give the ultimate power to the one person, or the group of people, that operates the blog. If third-party comments are enabled, they remain a secondary matter and are usually displayed less prominently, often not even showing up on default until a reader follows a "view comments" link. Even then, the blog owners have the power to remove specific comments or to even disable comments altogether. [5]

Blogs are, however, potentially much more than just plain logs. Posting comments or linking to other blogs facilitates the creation of online friendship networks between people who share common interests or discuss common topics. So, in a sense, blogs form online communities [1,3,4]. Apart from traditional blogging software there are the more recently created "social networking websites", which often include blogs as one of their many other features. Social networking websites focus on forming communities and online friendships. These usually have more developed features for interpersonal communication, such as: sending private messages between users, maintaining friendship lists and sharing photo or video collections. While some studies that will be mentioned here do go into the wider area of social networking, this review will focus more on the more specific subject of blogging.

A worldwide popularity of blogs as a new publishing and communications medium has created privacy problems for some bloggers, as their thoughts become available for anyone to read and search. This review will first discuss the problems that can occur due to privacy loss, then we will go over the ways in which people can be susceptible to privacy loss and we conclude with an overview of the discussions of various privacy protection techniques.

## PRIVACY LOSS

Snyder et al. [6] provides a table featuring some of the many blogging-related incidents that have made news headlines. There have been numerous cases of employees been fired for negative blogging about their workplace or employer; some bloggers even got into work trouble over writing about their private lives, such sexual experiences or racial thoughts. As mentioned in the same review, for

Chinese bloggers the consequences can be especially dire, there have been several incidents of bloggers being imprisoned for up to 10 years after publishing negative material against the authorities. In Bortree's study [1] of teenage girls' blogging habits, several girls reported that in the past they had been contacted by online stalkers who had made inappropriate comments and even threats; one girl stopped blogging altogether after such an incident. This is worrying as a Bortree cites 2003 statistics that showed that 51% of bloggers are teenagers and 56% are female, making teenage girls the largest demographic group of bloggers, at least at the time of the survey.

Apart from these extreme cases, more typically bloggers may run into social problems over their writing being revealed to those whom it was not intended for. For instance, putting friends in a negative light, and expecting them not to find out, can turn into nasty social situations if a simple web search reveals their writing to the person whom they were writing about [1,5]. Or employers, who cannot ask such questions in job interviews, could use information gathered from blogs to quietly discriminate against prospective employees on the basis of their private lives, political opinions, sexual orientation, religious beliefs, etc. [6]

#### **SUSCEPTABILITY TO PRIVACY RISKS**

Viégas [7] attributes most of the privacy flaws of blogs to their nature – while they are easy to use and available to anyone with an internet connection, the content posted is persistent. Though previous posts can be manually erased from a blog, they can stay online for years if left untended (even when a post does get deleted, duplicates of this information can stay on in other places on the Web – such as search engine cache or parts that have been quoted in other blogs).

Viégas' publication is based on reviews of previous work as well as an original questionnaire-based survey of bloggers. The questionnaire was advertised on university mailing lists as well as several high-profile blogs. Due to this self-selective bias<sup>1</sup> in this study the results of the questionnaire can only be used to identify issues that some groups of bloggers may face, they cannot be related to bloggers as a whole and it cannot be concluded that no other problems occur. Viégas' states that most bloggers understood the persistent nature of their content, however a majority still published personally identifiable information. Further, most wrote about others in their

---

<sup>1</sup> Only respondents who had sufficient desire to respond and had the time went to do the survey. Also, due to the places where the blog was advertised, the demographics of the respondents were strongly slanted towards American-based university students and staff (though some other demographics were also captured).

blogs and shared private experiences without permission, sometimes using real names of others. However those who did happen to strike problems over a carefree attitude in the past were much more careful. Viégas expressed some concern that many bloggers seemed unaware that they are legally liable for what they write, as they would be in any other medium the laws of libel apply. A common problem area is blogging about employment, where Viégas suggests that more employers create guidelines on what is acceptable or unacceptable for their employees to discuss in their blogs, as some companies have done (two famous examples being Google and Microsoft).

Nardi et al. [5] carried out a similar study to that of Viégas. They chose blogs of Stanford University staff and students, as well as some that were linked to by the university-related blogs. Analysis of blog content was performed, later followed by face-to-face, email and phone interviews. Similarly, this study's results are not indicative of the wider population of bloggers as a whole due to statistical bias; however it still provides an insight into the views of some bloggers. As with Viégas, this study also indicated that many bloggers tended to be not so concerned about their privacy and some gave out accurate contact information with their real names. While some respondents wanted a wide readership and attention, others seemed to be more indifferent, and thought that no one but their close friends and family would be interested in their blogs, though understanding that anyone could access their writing. Similarly, in Bortree's study of teenage girls' blogs [1] most blogged to keep up with their friends and had no other target audience in mind, many girls not expecting that strangers, nor older people, would read their blogs. Some girls specifically said that they did not want anyone but their friends to read their blogs.

Frankowski et al. [2] explores the issue of *re-identification*, which occurs when someone's pseudonymous username from one website can be linked to a non-pseudonymous identity through matching of information that is replicated in both places, thus revealing their real-world identity. Even the minimal amount of information can be used to extract someone's identity. For example 87% of Americans from the 1990 census could be identified uniquely by just their date of birth, gender and their ZIP code. The problem with this is that while someone may be willing to reveal their name on the Web in one kind of discussion, they may not want to be known in the other environment (say if they were discussing workplace problems). To provide an example of how this can be performed automatically, the study performs re-identification to match users on a film discussion forum to users on a film ratings website. Out of all the users who had mentioned 8 or more films in forum discussions, 60% uniquely matched to accounts on the ratings website. However the accuracy of this re-

identification is only assumed and is not analysed in the study.

Gross et al. [3] undertook a study of privacy risks on Facebook<sup>2</sup>, a social networking website for university students. In the sample consisting of Carnegie Mellon University (USA) students 21.2% of males and 15.4% of females revealed enough information about their location and daily routine that they could be susceptible to real life stalking, while 44.3% of all students were susceptible to demographic re-identification; as also mentioned by Frankowski et al. – with the date of birth, ZIP code and gender available. Further, those of the students who tried to keep their names anonymous but posted photos a high number was successfully matched to fully identified photos hosted on their university's websites, using a commercial facial recognition tool. Another worrying threat that was pointed out was identity theft – American Social Security Numbers are can be “estimated” if someone's phone number, ZIP code, place of birth and date of birth are available (most of the digits of Social Security Numbers are based on these pieces of information). As Social Security Numbers are widely used as identification throughout USA, someone's identity could be impersonated. Facebook presents a somewhat more secure case than typical blogs – only registered users can read account profiles, and only those with official university email addresses can register. However this is more of a deterrent rather than a stringent security measure – the study points out several different methods in which this perceived security can be circumvented both by registered users from other universities and by people who are not university students at all.

#### **PRIVACY PROTECTION METHODS**

Most blogging hosts allow their users to control who sees their blog or even specific posts on their blog. This control can be through methods such as password protection or requiring readers to first register an account and then be added to the author's “friends” list. Kozlov [4] reviews such privacy controls of LiveJournal<sup>3</sup>, he gives a positive light to the fact that they are not treated as a fixed structure but rather a “privacy management suite” which evolves with the feedback from the blogging community. LiveJournal lets people maintain “friend” list, and have three levels of read protection “public”, “friends only” and “private”, writing rights to someone's blog can also be shared – allowing some friends to share blogs.

---

<sup>2</sup> Facebook – <http://www.facebook.com/>

<sup>3</sup> LiveJournal (<http://www.livejournal.com/>) is a popular blog host, it is described by Kozlov [4] as a “hyper-blogging” service, which he defines as blogging with extra inter-user communication features.

However, the studies of Nardi et al. [5] and Viégas [7] concluded that controls like these are utilised only by a small number of bloggers – almost all bloggers in their studies kept all their posts accessible to anyone even if such protection features are available. On the contrary, Kozlov seems to imply that these features are actually used frequently at LiveJournal with wide user satisfaction – this could be due to LiveJournal's different community structure and blogger demographics from typical blog hosts.

Snyder et al. [6] discusses the Terms of Use policy of MySpace<sup>4</sup> – which states that “using any information obtained from the MySpace Services in order to harass, abuse or harm another person MySpace”. Technically, this policy prohibits people from using MySpace to discriminate against MySpace users based on their beliefs. However, the same study quotes an American teenager being arrested for vandalism after he put pictures of himself spray-painting a church on his MySpace account, and three American policemen being suspended for their derogatory comments against homosexuals posted on MySpace.

Of course none of these tools keep identifiable information such as email and IP addresses from the actual blog host – which can create problems if the local authorities want to find out the identity of a blogger (such as the previously-mentioned situations with the Chinese government). Viégas [7] mentions Invisiblog, a now-defunct host which tried to tend to this issue by not storing any information about the bloggers – not even email or IP addresses. However, neither the adequacy of such protection, nor the legal issues of such a service were discussed.

Frankowski et al. [2] analysed two possible methods that users could use to prevent re-identification with their algorithm for movie forums and movie ratings – suppression and misdirection. Suppression involves intentional failure to mention some key information which may be used to link two different online identities together, while misdirection involves intentionally mentioning information which is either incorrect or cannot be linked to the other identity. Suppression produced favourable results only when large ratios of data were suppressed, misdirection produced more favourable results and showed to be more efficient when mentioning popular items.

#### **CONCLUSION**

Currently the privacy protection features are prominently available on the technological level (access controls) and sometimes on the contractual level (Terms of Use

---

<sup>4</sup> MySpace (<http://www.myspace.com/>) is a social networking website that includes blogging features.

limitations). These solutions are not always effective as problems still occur. A large part of it is due to under-utilisation of technology by the bloggers themselves, as well as their occasional carelessness about the information they publish. Most are indeed aware of the nature of the medium in which they publish the information, but are not so aware of the consequences that they may incur and hence are not concerned about neither what they say nor about setting limits on who may read it. However, those who do find problems end up regretting taking the issue lightly.

Automated re-identification is a worrying threat. Internet access is becoming faster, more widely available and more information is becoming available on the Web. More advanced analysis tools could become more readily available to the average users with time, so perhaps in the future such data-mining and facial recognition techniques (as in [2,3]) could be performed by almost anyone. As the introduction of search engines in the 1990s had, for many, unexpectedly revealed private websites and USENET discussions [7], automated re-identification could reveal information which was not visible on the superficial level.

From what has been seen here, the most fail-safe way to keep privacy on the Web seems to be just to avoid publishing any information that could potentially be personally identifiable anywhere on the web, or to limit it to areas where access to it is strictly limited. Of course this can be an unreasonable inconvenience to most bloggers.

#### FURTHER RESEARCH

Though blog analysis studies and blogger surveys have been carried out in [1,3,5,7], all of these had strong selection bias (limited to: specific network of friends, self-invited respondents, limitations to certain US states only or to certain universities) – none attempted to have representative samples of bloggers as a whole. Therefore, while they show some common issues and behaviours, they cannot be used to describe all bloggers and may be giving undue weight to some issues while missing others. All these studies have focused on identifying problems, but there seems to be little discussion on potential solutions, or analysing the efficiency, utilisation and usability of current solutions.

Several other studies ([1,4,5,7]) have suggested that people behave differently with their online personas than they would if they were interacting with someone face-to-face, however these suggestions usually look more like the authors' assumptions and perceptions rather than something more solid stemming from statistical evidence. It would be interesting to see if this is really so and to which extent people's blogging personas differ from their physical selves.

Another important area that needs further investigation is the lack of perception of legal, and social, liabilities for content published on blogs. While technological solutions can reduce the risks, this lack of perception seems to be the actual root of most problems.

The studies by Frankowski et al. [2] and Gross et al. [3] analyse re-identification on some social websites, but not blogs specifically. Further work on re-identification could be done with respect to blogs, which is a more complicated issue – while social networking websites tend to have a fixed structure with certain data fields always present in an organised fashion, blogs are usually more open ended.

#### REFERENCES

1. D. S. Bortree. Presentation of self on the web: an ethnographic study of teenage girls' weblogs. *Education, Communication & Information*, 5(1):25–39, March 2005.  
<http://taylorandfrancis.metapress.com.ezproxy.auckland.ac.nz/content/g0t12g401552g44u/>
2. D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl. You are what you say: privacy risks of public mentions. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 565–572. ACM Press (2006).  
<http://doi.acm.org.ezproxy.auckland.ac.nz/10.1145/1148170.1148267>
3. R. Gross, A. Acquisti, and I. H. John Heinz. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005. ACM Press.  
<http://doi.acm.org.ezproxy.auckland.ac.nz/10.1145/1102199.1102214>
4. S. Kozlov. Achieving privacy in hyper-blogging communities: Privacy management for ambient intelligence. In *Proceedings of the WHOLES Workshop: a Multiple View of Individual Privacy in a Networked World*, pages 1–9, 2004.  
<http://www.sics.se/privacy/wholes2004/papers/kozlov.pdf>
5. B. A. Nardi, D. J. Schiano, and M. Gumbrecht. Blogging as social activity, or, would you let 900 million people read your diary? In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 222–231. ACM Press (2004).  
<http://doi.acm.org.ezproxy.auckland.ac.nz/10.1145/1031607.1031643>
6. J. Snyder, D. Carpenter, and G. J. Slauson. Myspace.com - a social networking site and social contract theory. In *Proc ISECON 2006*, volume 23,

pages 1–9. EDSIG, the Education Special Interest Group of AITP, 2006.

<http://isedj.org/isecon/2006/3333/ISECON.2006.Snyder.pdf>

7. F. B Viégas. Bloggers' expectations of privacy and accountability: An initial survey. *Journal of Computer-Mediated Communication*, 10(3), article 12, April 2005.

<http://jcmc.indiana.edu/vol10/issue3/viegas.html>