Forensics Software for Detecting Online Paedophile Activity

James Walkerdine[1], Phil Greenwood[1], Awais Rashid[1], Paul Rayson[1], Corrine May-Chahal[2]
[1]*Computing Department, Lancaster University, UK*
{walkerdi|greenwop|awais|paul}@comp.lancs.ac.uk
[2]*Applied Social Science, Lancaster University, UK*
c.may-chahal@lancaster.ac.uk
Penny Duquenoy, Patrick Watson
EIS, Middlesex University, UK
{p.duquenoy|p.watson}@mdx.ac.uk
Matt Jones
Computer Science, Swansea University, UK
mattjonez@gmail.com
Margaret Brennan
Applied Psychology, University College Cork, Ireland
maggiebrennan50@gmail.com

Abstract

Recent years have seen a rapid rise in the number and use of online social networks. Such social networks vary in nature from chat systems, for example, MSN, Skype and IRC, to online communities, such as, MySpace and YouTube, through to file sharing systems, for instance, peer-to-peer networks: Gnutella, BitTorrent, FastTrack, etc. Amongst the many types of 'risk' on the internet as mentioned in the Byron review in the UK and Internet Safety Technical Task Force in the US, these social networks pose two significant risks in terms of child exploitation.

The first major type of risk is paedophiles and other child sex offenders predating on children. Children actively participate in social interactions using forums such as chat rooms and web-based communities. Offenders can use such forums to predate on children, or even to plan the commission of sexual offences against children. These concerns are reflected by the formation of the Virtual Global Taskforce and specialist UK enforcement agencies and Scottish legislation to criminalise the 'grooming' of children in chat rooms in October 2004.

The second risk is the offence of distributing and sharing child abuse media. Child sex offenders can formulate their own social networks using mechanisms, such as file-sharing systems, in order to distribute and share child abuse media. The scale of distribution of illegal media (including child abuse media) on such file-sharing systems was highlighted by a recent study at Lancaster University, which found that 1.6% of searches and 2.4% of responses on the Gnutella peer-to-peer network relate to illegal sexual content. Given the system's scale, these results suggest that, on the Gnutella network alone, hundreds of searches for illegal images occur each second. The study also found that, of those users sharing illegal sexual content, 57% were solely devoted to such distribution while half of the material shared by another 17% involved such content.

The EPSRC funded Isis project (www.comp.lancs.ac.uk/isis/) aims to help tackle these risks by developing an ethics-centred monitoring framework and toolkit that will support law enforcement agencies in the task of policing online social networks for the purpose of protecting children. Specifically it seeks to tackle three significant research challenges:

1. *How to identify active child sex offenders across online communities?*
   Distinguishing between "innocent" interaction amongst children or children and adults, and the predatory advances of paedophiles and other child sex offenders who will often masquerade as a child or friend in order to gain a potential victims trust
2. *How to identify the core distributors of child abuse media?*
   Accurately identifying child abuse media from the plethora of material that exists within file sharing systems; media that offenders often describe using specialised vocabulary that evolves and changes over time.
3. *How to ensure that any technical solutions maintain ethical practices?*
   Balancing the benefits of using technology for child protection with the need to protect innocent users from the potential of falsely being identified as child sex offenders and safeguarding their privacy.

Given the vast amount of information that is communicated within online social networks, new monitoring and analysis technologies are required. Within Isis novel techniques are being developed that will support the analysis of chat logs and the non-invasive real-time monitoring of file sharing networks by drawing upon natural language processing practices. These can be used, for example, to help determine whether certain terminology repeatedly crops up, or whether users have trademark phrases that they use. Sophisticated statistical analysis techniques allow for the creation of language profiles for certain group types or even for an individual. These have already been successfully used to automatically distinguish between speech from males and females, from over 35 year olds and under 35 year olds, and also speakers from different regions within the UK. As part of Isis these techniques will be extended to allow the differentiation of child and adult language, and by extension the detection of adults masquerading as children. Law enforcement personnel will also be able to build language profiles of known paedophiles, which could then be used to assist in determining if they are re-offending. It should be noted, however, that our goal is not automation but to provide support for detecting potential sexual offences through analysis of large amounts of data which cannot manually be analysed in an efficient manner.

In parallel the Isis project is also studying the ethical and social implications of the technology that is being developed and, as a result, further informing its development. To date there has been a lack of suitable case studies in the computer ethics literature and appropriate guidance for technology developers to incorporate ethical considerations within the development cycle. Within Isis we are also developing new understandings of user-centred development methods for highly sensitive systems. These, in turn, will benefit future related developments and help mitigate the effects of adverse outcomes that impact on public acceptability.