Supporting Law Enforcement in Digital Communities through Natural Language Analysis

Danny Hughes[1], Paul Rayson[1], James Walkerdine[1], Kevin Lee[2], Phil Greenwood[1], Awais Rashid[1], Corinne May-Chahal[3], and Margaret Brennan[4]

[1] Computing, InfoLab 21, South Drive, Lancaster University, Lancaster, UK, LA1 4WA
{danny, paul, walkerdi, greenwop, marash}@comp.lancs.ac.uk
[2] Isis Forensics, P.O. Box 793, Lancaster, LA1 9ED, UK
k.lee@isis-forensics.com
[3] Department of Applied Social Science, Lancaster University, Lancaster, UK, LA1 4YL
c.may-chahal@lancs.ac.uk
[4] Child Exploitation and Online Protection Centre, 33 Vauxhall Bridge Road,
London, UK, SW1V 2WG
maggie.brennan@ceop.gov.uk

Abstract. Recent years have seen an explosion in the number and scale of digital communities (e.g. peer-to-peer file sharing systems, chat applications and social networking sites). Unfortunately, digital communities are host to significant criminal activity including copyright infringement, identity theft and child sexual abuse. Combating this growing level of crime is problematic due to the ever increasing scale of today's digital communities. This paper presents an approach to provide automated support for the detection of child sexual abuse related activities in digital communities. Specifically, we analyze the characteristics of child sexual abuse media distribution in P2P file sharing networks and carry out an exploratory study to show that corpus-based natural language analysis may be used to automate the detection of this activity. We then give an overview of how this approach can be extended to police chat and social networking communities.

Keywords: Social Networks; P2P; Network Monitoring; Natural Language Analysis; Child Protection.

1 Introduction

Social networking sites, chat applications and peer-to-peer (P2P) file sharing systems support millions of users and have redefined how people interact on the Internet. Millions of people use social networking sites such as MySpace [1] and chat applications such as MSN messenger [2] to support their personal and professional communications creating so-called *digital communities*. Similar communities have been created in P2P file sharing systems, such as Gnutella [3], which enables the decentralised sharing of files free from control or censorship by third parties.

As the scale of digital communities and their importance to society grows, there is a strong tendency for them to reflect real communities with the infiltration of criminal

S.N. Srihari and K. Franke (Eds.): IWCF 2008, LNCS 5158, pp. 122–134, 2008.

activities. For example, social networking sites have given rise to the phenomenon of 'cyber-stalking' [4] while chat applications have been used by paedophiles to support online victimisation, such as the 'grooming' of children [5]. Similarly, P2P file sharing systems have been implicated in both copyright infringement and the distribution of material related to child sexual abuse [6]. Digital communities, therefore, present two major and related challenges to law enforcement:

1. Digital communities provide new and easier ways for criminals to organise. An example of this is the ability of paedophiles to formulate ad-hoc networks to exchange child sexual abuse media and, even more seriously, to plan child sexual abuse activities.
2. The extremely large scale of digital communities coupled with rapidly evolving underlying protocols renders pro-active manual policing (analogous to police patrols in the physical world) infeasible. This can leave children and vulnerable adults exposed to significant risk.

This paper focuses on advancing methods for safeguarding children in digital communities and argues that effective pro-active policing of digital communities requires the use of automated language analysis techniques. These techniques, inspired by the fields of computational and corpus-based linguistics, can be used to help law enforcement agencies to identify:

i. Criminal activity.
ii. Evolving and emergent criminal terminology.
iii. Child predators masquerading as children.

To illustrate the scale of this problem, we perform an analysis of the scale and characteristics of the distribution of child sexual abuse media on P2P file sharing systems. Our study reuses tracing data collected and reported on in a previous study [6], however the analysis reported here is new and focuses specifically on child sexual abuse. We then describe an exploratory study involving the classification of emerging criminal terminology using natural language analysis. Finally, we discuss how this approach could be extended to support the policing of chat and social networking sites in order to prevent paedophiles using these systems to victimise children. While, in this paper, we focus on tackling paedophilia in digital communities, the techniques can be expanded to other criminal activities, for instance, to identify terrorists recruiting impressionable youth or organising attacks through digital communities.

The remainder of this paper is structured as follows: section 2 provides an overview of the prevalent digital communities of today. Section 3 introduces the background to our natural language analysis approach and its application to the policing of digital communities. Section 4 presents an analysis of the scale and characteristics of the distribution of child-abuse media on P2P file sharing systems. Section 5 evaluates the effectiveness of natural language analysis to automate the identification of child-abuse media. Section 6 provides a discussion on how natural language analysis can also be used to support the pro-active policing of chat and social networking sites for various criminal activities, and section 7 concludes and highlights areas of future work.

2 Digital Communities

Today's major digital communities include: P2P file sharing systems, chat applications and social networking sites. We first review the existing evidence of criminal activity in these communities and subsequently highlight the policing challenges.

2.1 P2P File Sharing

Peer-to-peer file sharing systems use the bandwidth and disk space of home or office computers to maintain a distributed online library of files. Any P2P user may directly share files with any other user, in a largely anonymous and censorship-resistant way. P2P has revolutionised how people use the Internet to communicate, empowering users to produce and distribute content free from any form of control or censorship by third parties. Today it is widely accepted that P2P applications (for example, Gnutella [3] and Bittorrent [7]) are responsible for the vast majority of internet traffic [8].

Unfortunately, this perceived anonymity has led to wide-spread illegal behaviour, most notably the use of P2P file sharing systems to distribute illegal content. Current research suggests that 90% of all material on P2P file sharing networks is copyrighted [9]. Likewise a recent study [6] showed that 1.6% of searches and 2.4% of search-responses on the Gnutella P2P network relate to illegal sexual content such as images of child sexual abuse. These levels equate to hundreds of searches for illegal images per second. While this work illustrates the significant scale of this problem, it also shows that the distributors of such material tend to form sub-communities, which do not interact with the broader P2P community.

2.2 Chat Applications

In a similar vein to P2P, chat applications have had a dramatic impact on how people communicate. Although chat software has existed for decades in guises such as IRC [10], it is only within the last ten years that chat applications have become adopted by mainstream Internet users. This is perhaps best illustrated by Instant Messenger applications such as MSN [2] and Skype [11] that are now common tools in both the home and workplace.

As chat applications have become popular, they have also become a means to support criminal activities. Their growing use by children and young people has given paedophiles a new means to target children (for example, [12]), or even to locate other paedophiles and plan paedophilia-related activities (for example, [13]).

2.3 Social Networking Sites

Social networking sites such as MySpace [1] and Facebook [17] are the newest form of digital community to be adopted by mainstream computer users, and like chat and P2P file sharing applications, participation is expanding quickly— in June 2007 "Social Media Today" reported that Facebook alone had 25 million users [18]. As of April 2008, Facebook claims to support 70 million users [34]. As with chat and P2P systems, the growth of this type of digital community has created opportunities for criminal behaviour. Users of social networking sites are particularly vulnerable as these sites allow individuals to rapidly expand their social network to include un-vetted strangers. Furthermore, the tendency of some users to post personal information such as their real-world address, school/job and telephone number allows cyber-criminals unprecedented access to potential victims. In particular, this has been exploited by cyber-stalkers [19], who use personal information along with the access that social networking sites provide to harass their victims (for example [20]). While cyber-stalking has been recognised as an offence that can be prosecuted under a range of existing legislation (such as the UK Protection from Harassment Act, 1997), difficulties arise from the variable range of legislation in countries across the world. Social networking sites are also used by paedophiles as another avenue through which to approach children. For example, paedophiles have been observed masquerading as children in order to initiate contact with their victims [21]. While there is a need to educate users of social networking sites about protecting their personal data, there is also an urgent need for effective policing of these communities to protect vulnerable users, for example, children.

2.4 Problems Inherent in Policing Digital Communities

Efforts to combat the concerns highlighted above have been reflected by the formation of the Virtual Global Taskforce (VGT) [14], the launch of organisations such as the Child Exploitation and Online Protection centre (CEOP) in the UK [15], and the introduction of legislation to criminalise the 'grooming' of children in chat rooms [16]. However, law enforcement agencies policing digital communities in order to detect, prevent and prosecute criminal activity on these systems face three major challenges:

- ***Dealing with large volumes of data*:** With millions of users, digital communities generate vast amounts of data, which makes manual analysis infeasible.
- ***Identifying criminal activity*:** Criminals such as paedophiles often develop their own vocabulary of terms to disguise their activities.
- ***Identifying masquerading users*:** In most digital communities, the creation of fake user identities is trivial, allowing criminals to evade detection (e.g. paedophiles posing as a children to contact their victims).

Critically, the very characteristic of these digital communities - that is, the lack of routine feedback such as body language, tone of voice or facial expressions [4], hampers policing and inevitably puts greater emphasis on language use, which can, in turn, be used to aid policing. Specifically, we propose that internet policing problems may be tackled more effectively through automated natural language analysis of traffic originating from digital communities.

Such an approach can identify likely criminal activity amongst the huge volumes of innocent interactions occurring in digital communities and furthermore, through the use of language profiling, make it possible to detect paedophiles pretending to be children and other masquerading users (such as cyber-stalkers using fake personas). A brief overview of our natural language analysis approach and its application to detecting and preventing crime in digital communities is provided next. For more details on the set of natural language tools and techniques we deploy for the purpose, interested readers are referred to [33].

3 Natural Language Analysis

Existing work on policing online social networks has focused primarily on the monitoring of chat and file sharing systems. Chat policing software for home use such as Spector Pro [29], Crisp [30] and SpyAgent [31] allow the logging of online conversations, but are restricted in that they need to be installed on the actual PC that is being policed. In terms of language monitoring capabilities, the existing chat policing software tools rely on human monitoring of logs or simple-minded word or phrase detection based on user-defined lists. Such techniques do not scale. Nor do they enable identification of adults masquerading as children or support pro-active policing.

Techniques do exist which make use of statistical methods from computational linguistics and corpus-based natural language processing to explore differences in language vocabulary and style related to the age of the speaker or writer. Keyword profiling [22], exploiting comparative word frequencies, has been used in the past to investigate the differences between spoken and written language [23], British and American English [24], and language change over time. Rayson [25] extended the keywords methodology to extract key grammatical categories and key domain concepts using tagged data in order to make it scalable. The existing methodologies draw on large bodies of naturally occurring language data known as corpora (sing. corpus). These techniques already have high accuracy and are robust across a number of domains (topics) and registers (spoken and written language) but have not been applied until now to uncover deliberate deception. The second relevant technique is that of authorship attribution [26]. The current methods [27] would allow a narrowing in focus from the text to the individual writer in order to generate a stylistic fingerprint for authors.

Our particular focus in this paper is file sharing systems where filenames and search terms reflect specialised vocabulary which changes over time. Using a frequency profiling technique, we can find popular terms within the search query corpus. Known terms such as high frequency words normally expected within a general language corpus (e.g. the, of, and, a, in) can be eliminated either manually or by using a list of ‘stop words’, a technique often used in information retrieval. From this list emergent unclassified terminology may be identified, and through association classified, allowing for the detection of emergent criminal terminology.

Section 4 first explores the scale and characteristics of the distribution of child sexual abuse related media on P2P file sharing networks. Section 5 then explores the use of natural language analysis to terminology emerging from this deviant sub-community.

4 Child Sexual Abuse Media Distribution on File Sharing Networks

Our previous research has shown that P2P file sharing networks, and specifically Gnutella, are a major vector for the distribution of illegal sexual material [6]. We now revisit our experiments on the Gnutella P2P file sharing network in order to specifically analyze the extent to which this system is used to distribute media related to child sexual abuse. Section 4.1 presents our experimental methodology and section 4.2 presents our results.

4.1 Experimental Methodology

Gnutella is an open protocol designed to support the discovery and transfer of files among its users. In technical terms, the Gnutella protocol builds an unstructured, decentralized network where peers are required to forward network maintenance and file discovery messages, and to share files on the network. File discovery on Gnutella uses two message types. 'QUERY' search messages that are broadcast on the network with an XML or plain text search term to discover files. Plain text/XML 'QUERYHIT' search-response messages are used to inform a searching peer that a matching file is available. Thus by connecting to the Gnutella network and intercepting and analyzing QUERY and QUERYHIT messages we can explore what files users are searching for, and offering respectively.

In our previous experiment to quantify the level of resource discovery traffic that relates to illegal sexual material [6] we gathered a one month trace of Gnutella traffic (between February 27th and March 27th 2005). Two independent reviewers then analysed 10,000 search and search-response messages from three separate days within our trace (5th, 12th and 19th March), classifying them as relating to either illegal sexual material or other material. Our approach was to classify messages as relating to illegal pornographic material if they could only be interpreted as referring textually to such material. We found that an average of 1.6% of searches and 2.4% of search responses contained references to such material.

We have since revisited this data in order to analyze the level of resource discovery traffic that relates specifically to *child sexual abuse*. Once again, two independent reviewers analyzed three samples of 10,000 search and search-response text messages from the 5th, 12th and 19th March 2005. The results of our experiments are provided in section 4.2. As with our previous experiments, the understanding that can be provided through an analysis of meta-data is limited – while our results do show that the level of resource discovery traffic relating to illegal sexual material, there is no way to know whether these searches were generated deliberately or in error. Nor is it possible to know whether a user performing such a search found and/or downloaded the corresponding material. Thus the analysis of resource discovery traffic provides a useful, but imperfect measure of the extent to which P2P file sharing networks are used to distribute material relating to child sexual abuse. Section 4.2 presents the results of our experiments.

4.2 Level of Child Sexual Abuse Related Resource Discovery Traffic on Gnutella

The results of the two independent reviewers' traffic classifications are shown in tables 1a and 1b:

Table 1a. Reviewer 1, Child-Abuse Related Resource Discovery Traffic

	5th March	12th March	19th March
Search Messages	0.9% (90)	1.55% (155)	0.81% (81)
Search Responses	1.59% (159)	1.83% (183)	1.25% (125)

Table 1b. Reviewer 2, Child-Abuse Related Resource Discovery Traffic

	5th March	12th March	19th March
Search Messages	0.90% (90)	1.4% (140)	0.86% (86)
Search Responses	1.59% (159)	1.9% (190)	1.38% (138)

It can be seen that there is a high degree of correlation between the independent reviewers' classifications:

- An average of 1.07% of the search message sample was classified as relating to child sexual abuse. The minimum value we observed was 0.81% on March 19th, rising to 1.55% on March 12th. The standard deviation between samples was 0.32%.
- An average of 1.59% of the search responses relate to child sexual abuse. The minimum value observed was 1.25% on March 19th, rising to 1.9% on March 12th. The standard deviation was 0.25%. The higher level of search responses that contain references to child sexual abuse arises because search responses list multiple files.

Given the large scale of the Gnutella network, this equates to thousands of child-abuse related searches and search responses per minute, suggesting that P2P file sharing systems and Gnutella specifically are a major vector for this material. Unfortunately, a major problem in identifying the presence of child-abuse related media is the use of domain-specific terminology by paedophiles. Our trace data was further analysed by domain experts for the presence of non-standard child abuse terminology. We found that, on average 53% of searches and 88% of responses used such language. Section 5 discusses and evaluates the use of natural language analysis to identify such terminology emerging from criminal sub-communities.

5 Identifying Child Sexual Abuse Media Using Natural Language Analysis

This section introduces a simple scheme to detect and classify emerging terminology related to the distribution of child sexual abuse media. Section 5.1 describes our experimental methodology. Section 5.2 reports on our results. Finally section 5.3 discusses how this process may be fully automated.

5.1 Experimental Methodology

These experiments were performed using our Gnutella trace data [6] and assess the viability of using *association* to classify previously unknown terminology relating to child sexual abuse media. The classification process was performed in a semi-automated fashion by 10 volunteers with no domain-specific knowledge (i.e. links to child-protection services, law-enforcement agencies or related criminal convictions). Our hypothesis is that by identifying non-standard search words that have a high popularity and providing context by showing complete search phrases that contain these words, we can significantly improve the accuracy of manual catagorisation. A complete list of the search words being categorized can be obtained by emailing the authors (due to their sensitive nature they are not reproduced here).

Firstly, our natural language analysis tool (Wmatrix [28]) was used to create a frequency-ranked list of the 1000 most popular search words. From this list, the top 5 non-standard language search words were extracted for non-child sexual abuse related material and the top 5 non-standard language words were extracted for child sexual abuse related material. Each volunteer in our test-group was then asked to classify these words as either (i) related to child sexual abuse or (ii) unrelated to child sexual abuse.

Following this ‘blind’ classification of terminology, each volunteer took part in a semi-automated process wherein association with complete search phrases was used to suggest the meaning of the unknown search terms. Specifically, each volunteer was presented with 10 complete search phrases containing the unknown term that were selected at random from our trace data. Each volunteer then reclassified every term based on its appearance with known (English) language words contained in the search phrases presented. Section 5.2 describes the results of this experiment and section 5.3 describes how the process may be further automated in order to improve efficiency and scalability.

5.2 Experimental Results

Table 2 shows the results of the initial blind catagorisation and subsequent catagorisation via association with complete search terms. It can be seen that blind catagorisation of child-abuse related media had an average success rate of only 42%, while classification through association more than doubled the success rate to 94%. Firstly, this shows that natural language analysis can provide significant benefits for identifying emergent terms. Secondly, it illustrates that the association with previously classified language is a promising approach to supporting the classification of new terminology.

Table 2. Use of Association to Improve Classification of Child Sexual Abuse Related Terms

	Anonymous Reviewer									
	1	2	3	4	5	6	7	8	9	10
Success of blind catagorisation (%).	40	40	20	60	60	0	40	40	80	40
Success of catagorisation with association (%).	80	100	100	100	100	80	100	100	100	80

The results of this experiment are shown graphically in Figure 1 below. It can be seen that the ability of test subjects to blindly classify child sexual abuse related terminology varied considerably - from 0% for reviewer 6 to 80% for reviewer 9. In all cases the use of association improved the subjects' ability to correctly classify terms – by an average of 52%.

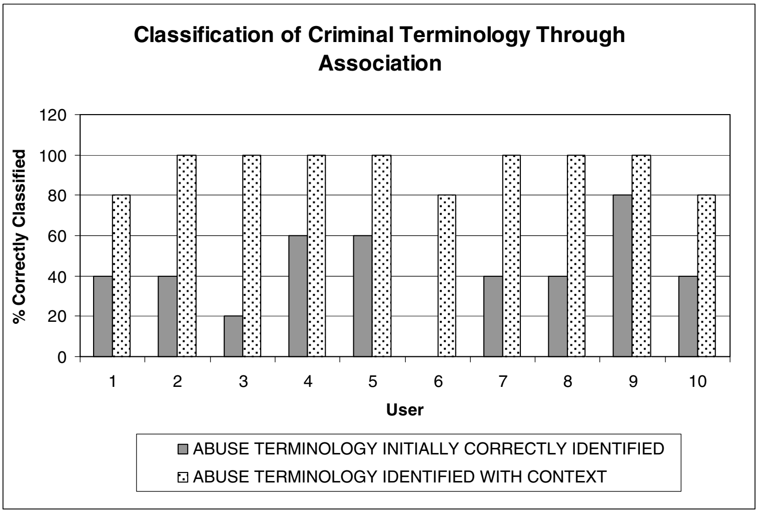

Fig. 1. Use of Association to Classify Child Sexual Abuse Related Terms

5.3 Automating the Detection and Classification of New Terminology

In order to automate the detection of new terminology, we can supplement the frequency profiling technique above with *comparative frequency analysis*. This technique compares the frequency profile generated from search terms to a frequency profile built from a reference corpus of general English. For each word in the profile, a log-likelihood (LL) statistical test is performed which estimates the significance of the difference in its frequency between the search term corpus and the reference corpus [32]. A larger LL value indicates that the difference in relative frequencies observed is larger and less likely to occur by chance. Therefore by sorting the resulting comparative profile on the LL value, we could find the most overused words in the search term corpus relative to an English standard reference. These key words appearing at the top of the list are most likely to be domain-specific words. This technique could be further extended by maintaining a *monitor corpus* of search terms consisting of trace data over the last 24 hours or 7 days, and using this as a reference corpus instead of the general English dataset. This would improve the extraction of novel terminology and enable the monitoring of short-term trends in search terms.

For classification of new terminology, we can exploit computer techniques that mirror the activities of the human volunteers during in the second stage of our xperiment. The extra information provided by viewing example terms in context provides evidence for the reader to make an educated guess at the meaning of an unknown term. A technique from corpus linguistics called *collocation* provides a way of calculating the strength of association between one word and others in the surrounding context. Pairs of words with high collocation scores occur significantly more often together than would be expected by chance based on their individual frequencies. Thus, if we find that an unknown word regularly collocates with other words that are known to be related to child-abuse media, we can make the assumption that the unknown word is also specific to that domain. This has the potential to more quickly identify emerging terminology and enhance the effectiveness of policing.

6 Discussion: Applying Natural Language Analysis to Chat and Social Networking Systems

The case for supporting the policing of chat and social networking systems is even more compelling than the case for supporting the policing of P2P file sharing systems. While P2P file sharing systems are a major vector for the distribution of existing media relating to child sexual abuse, chat and social networking systems are actively used to support the perpetrators of abuse. As such, supporting the pro-active policing of these systems is critical to preventing future crimes.

Fortunately, natural language analysis also holds significant promise for supporting the policing of chat and social networking systems. Domain-specific language relating to child abuse is largely unused except by those involved with the distribution of child abuse related materials. Thus, the detection of this language in chat rooms or on social-networking sites (e.g. its use by paedophiles planning a crime) may be used by law enforcement bodies to better target their limited manual policing resources.

In cases where paedophiles are masquerading as children on social networking or chat systems (e.g. to gain access to a child), the use of domain specific terminology is unlikely. In such cases, natural language analysis may still be employed to discover masquerading paedophiles. This requires the establishment and extension of corpora of child and adult language in chat rooms. Comparing samples of observed chat language with these corpora using the techniques described in section 3 may provide some clues to adults masquerading as children in terms of their vocabulary usage. However, we expect to have to extend these techniques to grammatical profiles in order to identify stylistic clues to deceptive language.

In either case the monitoring of digital communities also necessitates the development of new tracing functionality to support the establishment and extension of necessary corpora along with policing. For those digital communities which utilise peer-to-peer communication protocols, our existing monitoring approaches can be readily adapted [6]. However, those systems which use a more centralised communications infrastructure will necessitate the investigation of approaches similar to those used to implement web-based data mining [35].

7 Conclusions and Future Work

This paper has argued that natural language analysis is a promising approach to support the policing of digital communities. To illustrate the critical need for policing support, we first analysed the scale and characteristics of the distribution of child sexual abuse media on the Gnutella P2P file sharing network, finding that more than 1% of search messages and 1.5% of search-response messages relate to this material. We then showed that natural language analysis can be used in a semi-automated methodology to identify non-standard language used by those distributing child abuse media. Finally, we discussed how such an approach can be automated and applied across other digital communities such as chat and social networking systems.

The work presented here uses one of a number of techniques from corpus-based linguistics that we intend to trial in our approach to supporting the detection and classification of unknown vocabulary in chat and social networking systems. In future work we aim to: (a) to integrate the statistically sophisticated but knowledge-poor techniques from authorship attribution (that operate on one person's language) with linguistically-informed approaches in key domain analysis (that operate at the level of language of groups), and (b) to develop novel methods that are capable of detecting adults masquerading as children within the constraints of small amounts of language evidence, in the region of hundreds rather than thousands of words, that are observed in chat room data. Detection of unknown vocabulary is a relatively straightforward issue (comparable to a spell checker using a full form dictionary), but the task of classification of the unknown terms will require novel extension of techniques such as unsupervised word sense disambiguation where we will investigate exploitation of regular collocation with known domain vocabulary to seed the technique.

Any form of monitoring also raises ethical considerations. When extending and applying our monitoring approach to digital communities, such ethical issues will also be considered. This is important not only in respect of the abuse of children and the protection of the system developers, but also to the privacy of digital community members. The level of monitoring needs to be carefully and clearly explained to digital community members for them to accept the proposed approach. Part of the research performed will involve examining ethical issues and ensuring such issues are addressed throughout the development life-cycle of the monitoring approach.

Acknowledgements

This research is supported by a research grant from EPSRC/ESRC (reference EP/F035438/1) involving the Universities of Middlesex and Swansea as well as the Child Exploitation and Online Protection Centre (CEOP), project title "Isis: Protecting children in online social networks". We would also like to thank Isis Forensics (www.isis-forensics.com) for helping to conduct the experiments.

References

[1] MySpace (April 2008), `http://www.myspace.com/`

[2] MSN Messenger (April 2008), `http://webmessenger.msn.com/`

[3] The Gnutella Protocol Specification, version 0.4 (retrieved, April 2008), `http://www9.limewire.com/developer/gnutella_protocol0.4.pdf`

[4] Ellison, L.: Cyberstalking: Tackling Harassment on the Internet. In: 14th BILETA Conference: CYBERSPACE 1999: Crime, Criminal Justice and the Internet (1999)

[5] Pallister, D.: Internet paedophile gets nine years for sex with schoolgirls, Guardian Newspaper (June 23, 2006), http://www.guardian.co.uk/uk/2006/jun/23/ukcrime.davidpallister

[6] Hughes, D., Gibson, S., Walkerdine, J., Coulson, G.: Is Deviant Behaviour the Norm on P2P File Sharing Networks? IEEE Distributed Systems Online 7(2) (February 2006)

[7] Bittorrent Protocol Specification, version 1.0 (retrieved, April 2008), http://cs.ecs.baylor.edu/~donahoo/classes/5321/projects/bittorrent/BitTorrent%20Protocol%20Specification.doc

[8] Karagiannis, T., Broido, A., Brownlee, N., Faloutsos, M.: Is P2P Dying or Just Hiding? In: Proceedings of Globecom 2004, Dallas, Texas, USA (December 2004)

[9] Lee, K., Walkerdine, J., Hughes, D.: On the Penetration of Business Networks by P2P File Sharing. In: Proceedings of the 2nd International Conference on Internet Monitoring and Protection (ICIMP 2007), Santa Clara, California, USA (July 2007)

[10] RFC 1459: Internet Relay Chat (IRC) Protocol (retrieved, April 2008), http://www.irchelp.org/irchelp/rfc/

[11] Skype (April 2008), http://www.skype.com

[12] BBC News 24, Chat room Paedophile Jailed, http://news.bbc.co.uk/1/hi/england/2969020.stm

[13] BBC News 24, Men Jailed for Online Rape Plot (April 2008), http://news.bbc.co.uk/1/hi/england/6331517.stm

[14] The Virtual Global Task Force (April 2008), http://www.virtualglobaltaskforce.com/

[15] The UK Child Exploitation and Online Protection Centre (CEOP) (April 2008), http://www.ceop.gov.uk

[16] Scottish Parliament, The Protection of Children and Prevention of Sexual Offences (Scotland) Bill (April 2008), http://www.scottish.parliament.uk/business/bills/pdfs/b30s2-aspassed.pdf

[17] Facebook (April 2008), http://www.facebook.com

[18] Social Media Today, Facebook Explodes (June 2007), http://www.socialmediatoday.com/SMC/10670

[19] Office of Public Sector Information, Malicious Communications Act 1988 (April 2008), http://www.opsi.gov.uk/ACTS/acts1988/Ukpga_19880027_en_1.htm

[20] Crime Library (2007), Cyberstalking- A Case Study (April 2008), http://www.crimelibrary.com/criminal_mind/psychology/cyberstalking/5.html

[21] Panorama Transcript: One click from Danger (2008) (April 2008), http://news.bbc.co.uk/1/hi/programmes/panorama/7180769.stm

[22] Scott, M.: Focusing on the text and its key words. In: Burnard, L., McEnery, T. (eds.) Rethinking Language Pedagogy from a Corpus Perspective, Peter Lang, Frankfurt, pp. 104–121 (2000)

[23] Rayson, P., Leech, G., Hodges, M.: Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. Intl. Journal of Corpus Linguistics 2(1), 133–152 (1997)

[24] Hofland, K., Johansson, S.: Word frequencies in British and American English, NCCH, Bergen, Norway (1982)

[25] Rayson, P.: Matrix: A statistical method and software tool for linguistic analysis through corpus comparison, Ph.D. thesis, Lancaster University (2003)

[26] Holmes, D.I.: Authorship attribution, Computers and the humanities 28(2), 87–106 (1994)

[27] Juola, P., Sofko, J., Brennan, P.: A prototype for authorship attribution studies. Literary and Linguistic Computing 21, 169–178 (2006)

[28] Wmatrix (April 2008), `http://ucrel.lancs.ac.uk/wmatrix/`

[29] SpectorSoft 'Spector Pro' (April 2008), `http://www.spectorsoft.com`

[30] Protecting Each Other, Crisp (April 2008), `http://www.protectingeachother.com/`

[31] SpyTech 'Spy Agent' (April 2008), `http://www.spytech-web.com`

[32] Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, held in conjunction with ACL 2000, Hong Kong, October 1-8, pp. 1–6 (2000)

[33] Sawyer, P., Rayson, P., Cosh, K.: Shallow Knowledge as an Aid to Deep Understand-ing in Early Phase Requirements Engineering. IEEE Transactions on Software Engineering 31(11), 969–981 (2005)

[34] Face Book Press Information, `http://www.facebook.com/press/info.php?statistics`

[35] Peng, H.: A Data Mining Approach Based on Grey Prediction Model in Web Environment. Semantics, Knowledge and Grid, 76 (2006)