Technical report on

An Examination of Problematic Paraphilic use of Peer to Peer Facilities

Measurement and Analysis of P2P Activity Against Paedophile Content project
http://antipaedo.lip6.fr

Sean Hammond, Ethel Quayle, Jurek Kirakowski, Elaine O'Halloran, Freda Wynne

Abstract

This paper describes a methodology for investigating the paraphilic use of Peer to Peer facilities. The focus is upon problematic paraphilias, by which we mean those that imply illegal and/or non-consensual activity. The methods applied involve a new technique for evaluating the co-occurrence of paraphilic themes in order to inform a psychological profiling of P2P users. A typal analysis derived from Configural Frequency Analysis is reported. This shows in particular, that hebephilic and paedophilic behaviour are interrelated in a more complex manner than is expected by pure legal classification.

1. Background

Decentralised and anonymous P2P systems offer scope for the pursuit of socially dubious sexual interest in a relatively safe and secure environment. The ease with which pornographic materials can be accessed through P2P networks has raised serious concerns, particularly for the protection of children who may be recipients, or indeed the subjects, of such material (Congressional Committee on Government Reform, 2001; Greenfield, 2004). Nevertheless, the small amount of empirical research in the area suggests that pornographic exchange forms only a small part of the total P2P traffic. Thus, Hughes, Walkerdine, Coulson and Gibson (2006) found in a study of the Gnutella P2P network that pornography constituted only 1.6% of searches and 2.4 % or responses. This is in stark contrast to the warnings from US government agencies such as the US Federal Exchange Commission (2004), the US General Accounting Office (2003) and the US Congressional Committee on Government Reform (2001) on the pervasiveness of pornography on P2P networks. The CCGR (2001) demonstrates that the most popular Gnutella search terms in 2001 contained a number that were unequivocally sexual.

It should be said that Hughes et al (2006) were only concerned with 'illegal' sexual material and they used a very strict filter in order to reduce the number of false positives. This might suggest that their findings offer a conservative estimate of the sexual use of P2P networks. However, Hughes et al (2006) point out that those people using P2P networks to exchange pornographic materials, while representing a small sub-community of users, were particularly active. It must also be born in mind that even if a small proportion of exchanges are sexual in nature, the vast number of exchanges in general still suggests a very large number of sexually motivated P2P users. Hughes et al. do not profile the pornography users in their study but they do cite the importance of group-specific norms as a basis for the notion of a tightly identified sub-community of sexually motivated users. This leads to the intriguing possibility of a more specific analysis in which more tightly defined paraphilia based

sub-communities may be uncovered. The most troubling and contentious of these, of course, is paedophilia.

Much of the polemic against the P2P exchange of pornographic material is fuelled by the fact that such exchanges include material depicting the sexual abuse of children. The report from the General Accounting Office estimates that 42% of pornography exchanges on P2P networks involve children. Even allowing for fairly lax criteria for making such judgements, the anonymity and ease of access of P2P networks clearly facilitates the exchange of depictions of child abuse.

This paper describes a methodology for profiling paraphilic use of P2P networks. We use a specifically psychological-psychiatric focus to the profiling problem. The focus here is upon deviant patterns of sexual interest rather than the legal framework that defines sexual offending and abuse. For this reason we discriminate two paraphilic groupings associated with child molestation, these are the paedophile and the hebephile (Blanchard et al., 2008). While the paedophile manifests a sexual interest in prepubescent children, the hebephile is sexually interested in pubescent or recently post-pubescent children. Legally, the difference is largely unclear but psychologically the arousal pattern of the hebephile to secondary sexual characteristics is more in keeping with those of a normal adult. This does not suggest that hebephilia is less socially problematic than paedophilia but merely that there may be different underlying pathologies at work, so grouping them together may ultimately prove unhelpful.

1.1 Aims

The aims of this paper are to report on developments following an earlier study in which we explored the associations between sexual interest themes using exploratory and heuristic multidimensional scaling methods (Quayle, Hammond and Wynne, 2007). The aim of this report is to examine taxonomies of paraphilic sexual interest informed by the search terms individuals employ in their P2P interactions and it takes a more model-based approach to the examination of sexual interest themes.

2. Method

2.1 The Data

The present report is based upon two data sets made available through the MAPAP project team. The first of these data sets comprises a list of the 119,869 most commonly used terms in P2P submissions over the period of one week. This data set we will call the Search Terms List (STL). The second data set comprises all eDonkey transactions over this period and numbers over 3,000,000 records. This data set we will term the P2P Submissions List (PSL).

2.2 Identification of Thematic Categories

Our first priority was to try to impose some order upon the rich variety of the data collected. In the first instance, this involved an analysis of the STL in which the list was trawled for words with a sexual connotation. An exhaustive search of the list was carried out to identify terms that indicated sexually related material. In addition, a computer program was written for our Windows system in order to isolate words or part-words according to a given theme.

The result of this process was the identification of a number of specific themes defined by their sexual and fetishistic content. Specific terms and words associated with these categories were

identified from the STL data. It should be born in mind that these are not a final categorization an on-going refinement continues among the partners in MAPAP project. It should be clear that the motive behind using such search terms may vary and the assumption that these terms were being used to seek out sexually stimulating material is likely to be erroneous in some cases. However, given the large number of cases considered it was felt that this would constitute a manageable degree of error and would not invalidate the current methodology.

2.3 Identification of Individual's Sexual Interest Profiles

In order to provide a sexual interest profile for each record we next turned to the PSL data set. This contains over 3,000,000 records of submissions to eDonkey. A computer program was written for our Windows platform to scan these records in a serial fashion to find instances of the words identified as representing the thematic categories.

For each case, a record containing variables representing the 25 themes was created. Each variable was initially set to zero. If a sought after word occurred in the PSL data set for that case, then the variable representing the theme in which it is placed is incremented by one.

If, after scanning, a case has no occurrence of the critical words it is jettisoned and the program moves onto the next case. If, on the other hand, the case does contain critical words the record of 25 themes is written to a second data file. In this way the program identifies those cases where a user has made one or more sexually related submissions as defined by the terms recorded in appendix 1.

Thus a second data file was generated containing only these cases that manifest at least one of the 25 categories and this contained 62940 cases. Each of the 25 variables contains the frequency with which terms are submitted within each of the 25 thematic categories. In order to control for the fact that each theme is built of differing numbers of terms we chose to represent the data in binary form thus:-

$$\text{If } y_i > 0 \text{ then } x_i = 1$$
$$\text{If } y_i = 0 \text{ then } x_i = 0$$

Where yi is the observed frequency of words in thematic category i and xi is the binary value.

Each case, then, is recorded as a profile of 25 binary variables in which at least one variable is recorded as 1. Recording the data in this way provides a tractable data set to address the exploration of the relationships between sexual themes and a typal analysis of deviant sexual interest.

3. Results

The focus for this study is on problematic or 'deviant' sexual interests and to this end, a subset of the 25 themes was selected. These themes are presented in table 1 along with the percentage of total sexually motivated searches they each account for. A number of unexpected findings emerge. In stark contrast to the GAO claim that 42% of sexual exchanges on P2P networks involve children, we have found only 0.82% of searches to be explicitly paedophile orientated. It is also surprising to note that zoophilic or bestial interest appears to outstrip paedophilic searching.

The hebephilic theme had the largest incidence of these seven themes at 2.29% and the fact that this theme outnumbers the paedophile theme is not unexpected (Studer, 2004). All in all the

problematic themes selected here account for a relatively small amount of the 79427 sexually motivated searches observed in our study epoch (8.27%).

Theme	n	%
Gerontophilic	124	0.15%
Incest	455	0.57%
Paedophilic	657	0.82%
Bestial	907	1.14%
Sadistic	1046	1.31%
Rape	1561	1.96%
Hebephilic	1819	2.29%
Total Sexual	79427	

Table 1. Frequencies of Thematic Categories in the PSL Data Set

3.1 Configural Frequency Analysis

In order to examine the presence of distinct types in the data the search profiles using the seven paraphilic search themes were first subjected to a confirmatory zero-order Configural Frequency Analysis (CFA). The zero-order model was applied rather than the more typical first-order because the aim was to test the simple main effect of each paraphilic theme. In the zero-order model observed frequencies are tested against expected frequencies generated on the assumption of a uniform frequency distribution (von Eye 1990). This confirmatory analysis serves two purposes. First, it tests the hypothesis that each of the paraphilic themes represents a 'pure' and discrete sub-community in the P2P network space and second it serves as a tentative validation for the coding scheme used. The results are reported in table 2.

Theme	f	Lehmacher z	Adjusted p	Profile P H G B S R I
Incest	425	-3.02	p<0.0071	0 0 0 0 0 0 1
Rape	1377	40.07	p<0.0071	0 0 0 0 0 1 0
Sadistic	919	19.34	p<0.0071	0 0 0 0 1 0 0
Bestial	872	17.21	p<0.0071	0 0 0 1 0 0 0
Gerontophilic	123	-16.69	p<0.0071	0 0 1 0 0 0 0
Hebephilic	1763	57.55	p<0.0071	0 1 0 0 0 0 0
Paedophilic	615	5.58	p<0.0071	1 0 0 0 0 0 0
Bonferroni adjustment for p at 0.05 = 0.0071				

Table 2. Zero-Order Confirmatory CFA: Typal Identification of Paraphilic Interests

These findings support the contention that each of the themes describes a specific and independent sub-community. For example, of the 657 searches betraying paedophilic interest (table 1), 615 or 93.61% manifested a 'pure' profile with no other associated interest. It should also be noted that Incest and Gerontophilia are found to be significantly under-represented in the P2P network space,

as indicated by the negative z ratio in column 3. These profiles, where the observed frequency is less than the expected uniform frequency, are sometimes named anti-types to describe their scarcity in the sample under scrutiny (Krauth 1985).

As it stands this analysis provides limited information except that P2P users seeking paraphilic materials appear to be pretty specific in their searching behaviour. This does support Hughes at al.'s (2006) contention that the P2P network space is made up of distinct sub-communities providing the basis for targeted strategies for policing and managing problematic users. However, the question still remains as to whether there exists other 'comorbid' or multi-paraphilic 'communities'. A first order analysis will be required to test this hypothesis as the interaction between themes will need to be examined. For this analysis the expected frequencies are generated by conditioning out the main effects between themes. Such an analysis is summarised in table 3.

In this analysis all possible profiles or combinations of the 7 themes are included, making this an exploratory analysis. The number of possible profiles is 27 or 128, and it is instructive to observe that only 28 combinations are to be found in practice, suggesting a high degree of specificity in paraphilic searching behaviour.

In table 3 column 5 indicates whether each profile may be statistically identified as a 'Type' or an 'Anti-Type'. For present purposes we consider 'Types' to represent potential sub-communities in P2P network space. Note that the statistical criterion has changed because we now use a Bonferroni adjustment for the 128 potential profiles, allowing a more conservative estimate of significance.
In addition, a number of statistically significant profiles must be treated with caution because the expected frequencies are very small. This rules out profiles 36, 68, 98 and 100.

Only one combination type emerges and this is represented by profile 7 which includes rape and sadistic themes. It is perhaps not unexpected that such a combination would arise and it would seem to represent a sub-community of sexually sadistic individuals with an interest in rape. Of particular interest to us is the finding that combining paedophilia and hebephilia (profiles beginning 1,1,…) do not emerge in any significant manner. This suggests that the Child Molester label really does describe two distinct sub-groups in terms of sexual interest and may further suggest the development of discrete strategies for tackling both offender groups.

The anti-types are also revealing as they show the unlikely combinations. Thus hebephilic interest is particularly unlikely to covary with an interest in rape and sadism.

To conclude the typal analysis of this data a model-based approach was taken utilising a Latent Class Analysis (Lazarsfeld, 1950; Goodman, 1974; Magidson and Vermunt 2004). This is a probabilistic approach as opposed to the more deterministic CFA. The data, as described above, was truncated by removing all null profiles (0,0,0,0,0,0,0) and was then fitted to a number of unrestricted latent class models ranging from 2 to 8 underlying classes. The 7-class model was found to be the best fitting using the log-likelihood statistic and the Bayesian Information Criterion (BIC). This solution is summarised in table 4.

Profile	f	Expected	Lehmacher z		Profile P	H	G	B	S	R	I
1	56617	56642.76	-1.668		0	0	0	0	0	0	0
2	425	412.45	2.028		0	0	0	0	0	0	1
3	1377	1440.52	-6.098	A	0	0	0	0	0	1	0
4	3	10.49	-2.347		0	0	0	0	0	1	1
5	919	957.24	-4.277	A	0	0	0	0	1	0	0
7	119	24.34	19.565	T	0	0	0	0	1	1	0
9	872	828.17	5.202	T	0	0	0	1	0	0	0
11	24	21.06	0.652		0	0	0	1	0	1	0
13	1	14.00	-3.525	A	0	0	0	1	1	0	0
17	123	111.81	3.378	T	0	0	1	0	0	0	0
19	1	2.84	-1.107		0	0	1	0	0	1	0
33	1763	1685.70	7.055	T	0	1	0	0	0	0	0
34	6	12.27	-1.822		0	1	0	0	0	0	1
35	14	42.87	-4.525	A	0	1	0	0	0	1	0
36	7	0.31	11.976	T	0	1	0	0	0	1	1
37	3	28.49	-4.880	A	0	1	0	0	1	0	0
39	1	0.72	0.324		0	1	0	0	1	1	0
41	9	24.65	-3.217		0	1	0	1	0	0	0
65	615	597.49	2.392		1	0	0	0	0	0	0
66	9	4.35	2.247		1	0	0	0	0	0	1
67	11	15.20	-1.094		1	0	0	0	0	1	0
68	2	0.11	5.681	T	1	0	0	0	0	1	1
69	3	10.10	-2.262		1	0	0	0	1	0	0
73	1	8.74	-2.648		1	0	0	1	0	0	0
97	12	17.78	-1.397		1	1	0	0	0	0	0
98	2	0.13	5.200	T	1	1	0	0	0	0	1
99	1	0.45	0.815		1	1	0	0	0	1	0
100	1	0.00	17.369	T	1	1	0	0	0	1	1
Bonferroni adjustment for p at 0.05 = 0.00039											

Table 3. First Order Exploratory CFA: Paraphilia Profiling based on Thematic Interactions

Given the relatively large sample size, the statistical indices can be inflated so a more heuristic index of fit may be useful. The dissimilarity index is a descriptive measure indicating what proportion of the sample should be moved to another cell to get a perfect fit. On that basis the latent class solution reported suggests an excellent fit to the data.

Themes	Classes….. 1	2	3	4	5	6	7
Gerontophilic	0.000	0.998	0.000	0.000	0.000	0.000	0.000
Bestiality	0.008	0.028	0.026	0.000	1.000	0.019	0.054
Paedophilic	0.000	0.000	0.001	0.000	0.046	0.004	1.000
Hebephilic	0.000	0.035	0.008	1.000	0.010	0.020	0.003
Sadistic	0.000	0.031	0.001	0.000	0.000	1.000	0.000
Rape	0.000	0.000	1.000	0.001	0.000	0.000	0.001
Incest	0.997	0.000	0.000	0.000	0.000	0.000	0.000
Class Probabilities	0.019	0.072	0.143	0.279	0.230	0.101	0.154

Diagnostic Statistics

Log Likelihood	59.12
2LL	118.23
Pearson χ^2	194.60
Pearson χ^2 under independence	8281.38
Dissimilarity Index	0.004
BIC	93.90

Table 4. Latent Class Analysis of Paraphilia Themes: The 7-Class Solution

It is clear from these results that the best way to describe the underlying latent structure is as 7 discrete sub-communities each defined by one specific sexual theme. This is unsurprising given the CFA results above but it serves to further emphasise the contention of extant paraphilic sub-communities in the P2P network space.

4. Discussion

The research programme that supported this work is entitled 'Measurement and analysis of peer to peer activity against paedophile content'. The study reported here builds upon an earlier study exploring the paraphilic space indicated by P2P searches (Quayle et al 2007). In this study we take a further step in trying to derive a method for examining the latent groups of paraphilic users of P2P networks. As the title of the programme states our primary interest is in paedophile interest although this psychological/psychiatric term is often confounded with legal attempts to define sexual offenses against children. We have attempted to distinquish between offenders who may have a paedophilic orientation from those who may have a hebephilic interest. Not because one is of lesser concern than the other, but because if, as Hughes et al, (2006) suggest, P2P networks are constituted of specific sub-communities, an effective targeting strategy may prove to be a viable alternative to scattergun policing of the networks.

The findings presented here, are certainly suggestive of the existence of group specific social-norms, and while that cannot be directly ascertained with this data, there is ample evidence for specificity in the sexual interests expressed by people exhibiting paraphilic motivation for P2P use.

Acknowledgements

This work is supported in part by the MAPAP SIP-2006-PP-221003 project.

References

[1] Blanchard, R., Lykins, A. D., Wherrett, D., Kuban, M. E., Cantor, J. M., Blak, T., Dickey, R., & Klassen, P. E. (2008). Pedophilia, hebephilia, and the DSM–V. Archives of Sexual Behavior. 38, 335–350

[2] Congressional Committee on Government Reform (2001) Children's Access to Pornography Through Internet File-Sharing Programs. Special Investigations Division Committee on Government Reform. U.S. House of Representatives

[3] General Accounting Office (2003) File-sharing programs: Peer-to-Peer networks provide ready access to child pornography. GAO-03-351. Washington: US General Accounting Office

[4] Goodman, L.A. (1974) Explanatory latent structure analysis using both identi8able and unidenti8able models. Biometrika 61, 215–231.

[5] Greenfield, P.M. (2004) Inadvertent exposure to pornography on the Internet: Implications of peer-to-peer file-sharing networks for child development and families. Applied Developmental Psychology 25, 741–750

[6] Huges, D., Walkerdine, J., Coulson, G. and Gibson, S. (2006) Peer-to-Peer: Is deviant behaviour the norm on P2P file-sharing networks? IEE Distributed Systems Online. 7 (2), 1-11.

[7] Krauth, J. (1985) Typological personality research by Configural Frequency Analysis. Personality and Individual Differences. 6, 161-168.

[8] Lazarsfeld, P.F. (1950) The logical and mathematical foundation of latent structure analysis. In: Stouffer, S., et al. (Ed.), Measurement and Prediction. Wiley, New York.

[9] Lienert, G.A. (1988) Angewandte Konfigurationsfrequenzanalyse. Frankfurt: Athenaum.

[10] Magidson, J., and Vermunt, J.K, (2004) Latent class analysis. D. Kaplan (ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences, Chapter 10, 175-198. Thousand Oakes: Sage Publications.

[11] Quayle, E., Hammond, S., Wynne, F. (2008) An Empirical Investigation into the Sexual Interest Profiles Manifest in P2P Activity: A Preliminary Report. MAPAP Project. SIP-2006-PP-221005.

[12] Studer, L. H., Aylwin, A. S., Clelland, S. R., Reddon, J. R., & Frenzel, R. R. (2002). Primary erotic preference in a group of child molesters. International Journal of Law and Psychiatry, 25, 173–180.

[13] US Federal Exchange Commission (2004) Peer-to-Peer File-Sharing Technology:Consumer Protectionand Competition Issues: Staff Report. http://www.ftc.gov/reports/p2p05/050623p2prpt.pdf

[14] von Eye, A. (1990) Introduction to Configural Frequency Analysis. Cambridge: Cambridge University Press.

Project MAPAP SIP-2006-PP-221003.

http://antipaedo.lip6.fr

Supported in part by the European Union
through the *Safer Internet plus Programme*.

http://ec.europa.eu/saferinternet