Document Title:	RoundUp Predictive Tool (RPT) Project: Final Report
Author(s):	Marc Liberatore, Brian Neil Levine, Hanna Wallach, Janis Wolak, Thomas Kerle
Document No.:	248596
Date Received:	January 2015
Award Number:	2011-MC-CX-0001

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

RoundUp Predictive Tool (RPT) Project: Final Report
Grant No. 2011-MC-CX-0001
30 September 2014

Marc Liberatore (PI) (UMass Amherst), liberato@cs.umass.edu
Brian Neil Levine (UMass Amherst), brian@cs.umass.edu
Hanna Wallach (UMass Amherst), wallach@cs.umass.edu
Janis Wolak (Univ. New Hampshire), janis.wolak@unh.edu
Thomas Kerle (Fox Valley Tech. College), thomas.kerle@pobox.com

This report summarizes the results of the RPT project.

In Brief:

- We have characterized the trafficking of online CP trafficking across several peer-to-peer networks.
- We have distributed 17,567 surveys of complete CP cases, of which 12,621 (72%) have been completed and 3,532 (20%) refused. 1,414 (8%) remain pending. At least 1,227 (9.7%) of cases involving RoundUp were found to involve contact offenders.
- We have built and evaluated predictive models of contact offenders based upon the evaluation of peers' shared content. We have built and evaluated models to predict content type based upon file names.
- We have developed and deployed the latter models in production to the ICAC website.

Context of our project. Our project benefits synergistically from our leading of the design and deployment of the RoundUp suite of tools, as funded by other agencies. First released in 2009, the RoundUp tools are now standard in all 61 Internet Crime Against Children (ICAC) Task Forces across the U.S. for investigating p2p networks for trafficking in CP images. U.S. Federal agencies, the U.S. armed forces, and agencies in several other countries also use our tools daily. Over 7,000 investigators have been trained on our tools. Over 10,300 court-issued search warrants, investigating cases of possession and sharing of images of sexual child exploitation, have been executed based on the RoundUp project software. Most notably, since April 2012, investigators using RoundUp have identified over 850 past or current contact offenders sharing child pornography. Since April 2012, over 230 children have been rescued from sexually abusive situations due to investigations based on RoundUp software. All our tools are available without cost to law enforcement.

Detailed Report:

We have characterized the trafficking of online CP trafficking across several peer-to-peer networks. (Goal 1)

In two peer-reviewed publications we have analyzed the characteristics of CP files on p2p networks.

In "Measurement and Analysis of Child Pornography Trafficking on P2P Networks"[1] we evaluate the effectiveness of content various removal strategies, compare the aggressiveness of different sub-groups of peers, and look for evidence of attempts to evade detection. This paper is aimed primarily at a computer science reader.

In "Measuring a year of child pornography trafficking by U.S. computers on a peer-to-peer network"[2], we describe the functionality of p2p networks, how CP is exchanged on them, and how law enforcement responds. We characterize user behavior and file distributions across the Gnutella network in finer detail in this paper, aimed primarily at a social science reader.

PDF of both papers are included with this report.

[1] http://people.cs.umass.edu/~brian/bibliography/index.php?q=Hurley:2013; extended technical report available at http://web.cs.umass.edu/publication/docs/2013/UM-CS-2013-007.pdf

[2] http://dx.doi.org/10.1016/j.chiabu.2013.10.018

We have distributed 17,567 surveys of complete CP cases, of which 12,621 (72%) have been completed and 3,532 (20%) refused. 1,414 (8%) remain pending. At least 1,227 (9.7%) of cases involving RoundUp were found to involve contact offenders. (Goal 2)

For the first group of surveys distributed, covering cases opened before July of 2011, over 99% of 6,857 eligible surveys have completed or been refused, with less than 1% still pending. 8% of the completed surveys reported discovery of a contact offender. This and other references to percentage of contact offenders is across all peer-to-peer networks unless otherwise specified; there is some variance across networks.

For the second group of surveys distributed, covering cases from July of 2011 through September of 2012, over 99% of 5,106 eligible surveys have been completed or refused, with less than 1% still pending. 10% of the completed surveys reported discovery of a contact offender.

For the third group of surveys distributed, covering cases from September 2012 through July 2013, about 75% of 5,604 eligible surveys have been completed or refused (note: in the last semiannual the attached Survey Report for this group was correct, but the summary was incorrect), with about 25% still pending. 12% of the completed surveys reported discovery of a contact offender.

The proportion of contact offenders varied significantly per network. For example, the rate of contact offense in cases involving only Gnutella varies from around 6% to around 7.2%, depending upon the timeframe. The rate of contact offense in cases involving eMule is higher, around 14.8% over all the returned surveys.

Specific details of the current survey status are in the attachment named “2014-08 Survey Report.pdf”.

We have built and evaluated predictive models of contact offenders based upon the evaluation of peers’ shared content. We have built and evaluated models to predict content type based upon file names. (Goal 3)

Current tools (such as RoundUp:Gnutella) present investigators with the ability to search for specific terms of interest, to “browse” remote peers’ shared files, and to download files from remote peers. Investigators choose their targets for investigation on the basis of these data, their experience and intuition, and whatever policies their agency dictates.

Across all survey data, through June of 2014, procedures are identifying contact offenders in about one in ten cases, specifically, where the investigator indicated one or more of: an offender with prior arrest(s) for sexual offenses against minors; an offender charged with a previously unknown offense; and/or a case that raised strong suspicions, but where suspect was not charged (e.g., charges were barred by statute of limitations). In the remaining cases, the offender was either not a contact offender, or the investigator did not report a strong suspicion thereof. These numbers are consistent with previously reported data.

Models based upon Gnutella peers’ shared content are able to identify contact offenders at a rate about double background rate. On data from the first and second round of surveys, we achieved positive predictive value of about 11%

We found that the predictive models worked best (that is, had highest positive predictive value) when provided with many files upon which to base their predictions. Law enforcement prioritizes peers with many files for investigation (and subsequently removed from the network). Further, the resources available to do so appear sufficient to remove all peers with enough files to be usefully classified by our technique.

At the suggestion of our contacts and colleagues at ICAC and DOJ, we extended this same modeling technique to predict the type of file, based upon its name. Notably, the core of the prediction technique and the implementing code are nearly identical to that used to predict contact offenders. Here we met with considerable success.

As shown in the attachment “Content Type Prediction.pdf” , some categories can be predicted with extremely high precision (positive predictive value) and recall (sensitivity). We used a family of related techniques, and all do well. Almost 90% of the data set consists of “Age Difficult” (which was described to us as legally CP, but harder to prosecute as victims may not be pre-pubescent); prediction on this most common class is straightforward. Regardless, our classifier is able to correctly classify (precision) many instances (recall) of the other categories as well.

These results are useful to investigators, in that they provide a quick and relatively accurate first pass at identifying file types when the target of an investigation has hundreds or thousands of files, or when assessing many possible targets of investigation that each have few known files. We intend to continue this work in the future, characterizing content on other peer-to-peer networks of interest to law enforcement.

We have developed and deployed server-side models. (Goal 4)

As noted in previous reports, we have implemented our classifier in a portable and deployment-ready way. In cooperation with ICAC, we have deployed the content type prediction tool on the current ICAC server. We have also deployed it on the next generation production server, a newly-coded system and improved hardware, which is planned to be made live soon.

List of Attachments

1. hurley.www.2013.pdf
2. wolak.can.2013.pdf
3. 2014-08 Survey Status.pdf
4. Content Type Prediction.pdf