Internet Safety Technical Task Force
Technology Submission – Data Stream Profiling Tool

Appen Speech and Language Technology Inc
http://www.appen.com.au

JULY 21, 2008

ABSTRACT

Appen's Data Stream Profiling tool (DSP) was developed under US government funding and sponsorship to meet identified needs of intelligence and law-enforcement organizations. The DSP uses biometric modeling in order to identify individuals using remotely monitored computer terminals. By using a model based on typing behavior ("keystroke signature"), the DSP tool can attribute an incoming stream of keystroke data to an individual in its database. This capability can be applied to the problem of tracking the internet use of a person-of-interest (POI), such as a suspected internet predator, when they are known to use multiple terminals for access.

Keywords
verification, identification, profiling.

Functional Goals

Please indicate the functional goals of the submitted technology by checking the relevant box(es):

☐ Limit harmful contact between adults and minors
☐ Limit harmful contact between minors
☐ Limit/prevent minors from accessing inappropriate content on the Internet
☐ Limit/prevent minors from creating inappropriate content on the Internet
☐ Limit the availability of illegal content on the Internet
☐ Prevent minors from accessing particular sites without parental consent
☐ Prevent harassment, unwanted solicitation, and bullying of minors on the Internet
☒ Detection of activity of person-of-interest

PROBLEM INTRODUCTION

When a suspected internet predator (or other person-of-interest; POI) is identified, there are a number of options that can be used to monitor their internet usage. However, many of these rely on the person using a known computer, such as their home machine. If a POI chooses to use other internet connections, such as public internet cafes, monitoring becomes more difficult, and requires physical surveillance. Appen's Data Stream Profiling tool (DSP) can be employed for remote surveillance of computer usage, reducing this labor intensive physical requirement.

The DSP tool uses biometric modeling, specifically mathematical abstractions of a users typing behavior, in order to identify them. By gathering data from known periods of usage, such models can be created of any individual. This provides the database against which incoming streams of data from monitored locations are tested. Should a match be found, the DSP tool can generate an alert to inform an investigator of a POI's whereabouts. This process is akin to having automatic facial recognition on CCTV footage taken at areas of greatest risk from predators - near schools or at shopping precincts/malls for example. These approaches are being able to track individuals when they get into situations in which they may put others at risk.

PROPOSED SOLUTION

The operation of the DSP tool is based upon biometric models of an individuals typing behavior. Factors that contribute to such a model include among others: typing cadence; duration for which keys are held down; timing transitions between key sequences.

The DSP tool has three main components:

1. the Keylogger – the small software component that is installed (covertly if necessary) on any computer to be monitored. This is the component responsible to gathering the biometric data of interest and transmitting it, via standard internet connection, to:

2. the Matching Engine – the main component of the tool which generates biometric models and compares streams against these models. The engine can be configured by, monitored via, and transmits alerts through:

3. the Alert Control Panel – the interface to the DSP tool. The interface can be used by technical administrators to configure the operation of the matching engine, and to manage storage of streams and creation of models. It can also be used by investigators to configure alert generation, for example alert Agent Alpha when POI #231 uses any monitored terminal.

Models are generated from periods on known usage. This can be either from an individuals personal machine/login or from a period when a subject has been physically observed using a terminal at a recorded time. Models are created from multiple sessions and once created, physical surveillance is no longer required.

Sites that are monitored will transmit data to the matching engine, the frequency being configured by specific time or number of keystrokes. All incoming streams are compared against the database to see if they have been generated by a POI for whom there is a known keystroke signature. If this should occur, an alert can be generated to trigger an appropriate action. In the case of remote monitoring on an internet café, and alert will inform an investigator of a POI's activity, and they can then issue an order for retrieval of that monitored machines memory, in order to investigate further.

Additional Information:

Technical attributes – features and functionality.

Users can perform three main tasks with the DSP:

> Verification – validate that a person is who they claim to be.
> Identification – attribute a data stream to a known person.
> Profiling – determine the likely characteristics of a person from features of the data stream.

Solutions/Limitations

The DSP tool currently operates using a Keylogger component, requiring less that 10K disk space. A full operational version would be optimized to make it undetectable. It currently encodes keystream data as a mathematical abstraction, which means for privacy concerns, the real data that was typed (personal emails, passwords, etc) can be, but need not be, directly stored.

Hardware/Software requirements

No special requirements of machines to be placed under surveillance.

Application host machine is server based. Currently configured for Linux based system administration.

End-user aptitude

Investigator/analyst level – Low level computer skills only, would be required. The tool uses a simple web interface, and it was a requirement of the development sponsors that the interface be user friendly and oriented to the skill set of a main-stream law enforcement officer (i.e. not to the higher skill levels of officers working in a high tech crime environment).

System administrator level – advanced functionality requires system administrator level skills and may require some training.

Standards

The DSP and its contexts of use are novel, and as far as Appen is aware there are no relevant standards, existing or planned.

Reliance upon, and use of law and policy

The keylogger component of the DSP must be installed on any machines to be monitored. This may require a court order. There may also be privacy concerns if a public internet café is being monitored. Identification can be carried out on data at various levels of abstraction, therefore what was actually typed need not be inaccessible.

Viability of the technology in the US and Internationally

Though initial studies were conducted on data generated while typing in English, the vast majority of features are strictly biometric and therefore completely language independent.

APPEN OVERVIEW

Appen develops and markets sophisticated computer-based speech and language technology products and services for major international information and communication companies and government organizations. Appen is recognized as a global leader in the quality and quantity of its products and services.

Appen's products cover a range of text applications such as Natural Language Processing ("NLP"), as well as speech recognition, text-to-speech (speech synthesis), phonetic search, machine translation applications. These are developed from the high-end fusion of Appen's computer science and linguistics capabilities. Appen works in over 80 languages. It is headquartered in Sydney, Australia, with a US subsidiary Appen Speech & Language Technology Inc.

Appen has been operating for 12 years and is privately owned. The company has shown consistent growth since its creation. It is profitable and has a strong balance sheet.

Major customers are most of the world's large ICT companies, including Microsoft, IBM, Google, Motorola, Siemens, Toshiba and Nokia. Appen is also active in the US government sector in the defense and homeland security sectors. Appen has a number of strategic relationships with its major clients but is independent of any other organization.

Appen maintains an active R&D program with a growing portfolio of IP and patents. Developmental work includes:

- Tools for processing speech materials, especially to support the development of speech recognition and speech synthesis systems, and tools such as lemmatisers, tagging tools & spelling standardization or less commonly taught languages.
- Tools for content extraction from text materials, in particular the extraction of entity related information inn nominated categories
- Licensable software products for use by government intelligence and law enforcement agencies for forensic authorship profiling and tracking of persons of interest on-line.

Appen has a professional staff of approximately 50 people, most with post-graduate qualifications in computer science and/or linguistics. Key development people include

Dr. Julie Vonwiller (Project Director and Appen founder)
Dr Julie Vonwiller is a Director and co-owner of Appen. She has worked in the field of speech science and technology for more than 15 years.

Dr. Vonwiller has worked both in private industry and research, and has skills in Project Management of software-oriented projects, successful commercialization of R&D, management of multi-disciplinary research and development teams and in-depth familiarity with leading edge speech technology techniques. Her experiences ranges widely across lexical and text processing in many languages.

Julie has led several successful development projects for the US Government, including the Text Attribution Tool software development and the Data Stream Profiling software development.

Dr Scott Nowson (Computational Linguist)
Dr Scott Nowson is an experienced Computational Linguist with deep experience in the linguistic study of online diaries and weblogs. His research experience encompasses personality theory, language analysis tools, language with personality and gender, and language in computer mediated communication. Scott has extensive experience with the collection and processing of text corpora and sociobiographic data, and with data annotation for machine learning.

Scott was educated at the University of Edinburgh, where he completed a B.Sc. degree with Honors in 1999. Subsequently he was awarded a M.Sc. in Cognitive Science (2001) and a Ph.D. in Informatics (2005), also from Edinburgh University. Prior to joining Appen in 2007, Scott worked as a Research Fellow at Macquarie University investigating the application of natural language technology within the financial domains.

Stephen Norris (Software Development Manager)
Stephen Norris has provided IT consulting and contracting services to Appen since the company was first established. In June 2005, Stephen joined Appen as the IT group manager.

Stephen has extensive experience in software development and IT management, as well as being a highly skilled programmer and system designer himself. He has worked on projects ranging from high-performance, high-availability financial information systems through web-based stock trading, banking and transaction processing systems to document processing and management systems used for on-line billing and related systems.

Dr. Judith Bishop (Linguist)
Judith Bishop has worked as a Linguist and Project Manager in Appen since December 2004. Her experience in Appen has included research in phonetic analysis of speech, project management of large scale speech databases and language analysis in unusual languages. Judith has been a member of the Appen team working on the Text Attribution Tool development for the US Government.

Judith was awarded a PhD in Linguistics from the University of Melbourne (Australia) and completed graduate studies at Washington University in St Louis (USA) and University of Cambridge (UK).

Phil Hall (Advisor – Forensic Linguist)
Phil has a BA (Honors First Class) in Linguistics and the University Medal from Macquarie University and has received a number of scholarships, including a Scholarship from the Linguistic Society of America. He is completing a PhD in Linguistics focusing on forensic linguistics from Macquarie University. Phil has taught at university level and worked in a management capacity before joining Appen. Phil has acted as a consultant forensic linguist for the New South Wales Police in a number of extortion and stalking cases.

BUSINESS MODEL OVERVIEW

The Appen technology referred to in this submission can be made available under different business models. Typical users are envisaged to be

- Law enforcement organizations (federal, state and local
- Government agencies, especially in intelligence processing
- Relatively large commercial organizations who wish to provide screening in regard to their on-line user populations envisaged to be

Typical business models will be:

a) A system license in the form of a Software Development Kit (SDK) which will allow the system operator to incorporate the Appen technology into its own systems
b) A usage licensing agreement, with Appen providing the necessary servers and support, and users paying according to their level of usage.

Appen is in discussion with partners in the US market in relation to establishing all necessary support for deployment for the technology.

The technology has been delivered to the US Government as working prototypes, and is now undergoing final development ready for commercial deployment. Specific pricing has not yet been released.

MORE INFORMATION

Feel free to provide a URL that readers can go to for more information. This may include videos, detailed specs, or anything else that might be relevant. Indicate in this document what the readers might find if they go to the URL. This is a great place for information you would like to include that does not otherwise fit the structure of this document.

http://www.appen.com.au/files/products/DSP.pdf

CONTACT INFORMATION

Contact: Phil Hall (phall@appen.com.au)
Telephone: +61 2 9468 6321
Nth Tower, Level 6, 1 Railway Street
Chatswood, NSW 2067, Australia

CERTIFICATION

I certify that I have read and agree to the terms of the Internet Safety Technical Task Force Intellectual Property Policy.”