Analyzing Users' Web Surfing Patterns to Trace Terrorists and Criminals

Gabi Kedma, Mordehai Guri, Tom Sela, Yuval Elovici
Department of Information Systems Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel
gabik@post.bgu.ac.il, gurim@post.bgu.ac.il, tomse@post.bgu.ac.il, elovici@bgu.ac.il

Abstract— **Regular users, as well as criminals and terrorists, are using the Internet for various purposes. Today, current Big-Data technology allows law enforcement authorities to create a huge repository that contains all the surfing activity performed by all users over a considerable period of time. Our theoretical exploration suggests that such a repository contains valuable forensic information that may help to pinpoint suspected criminals or terrorists, and in certain cases can even preempt an upcoming criminal act. In this paper, we suggest how Big-Data analytics should be employed to perform such detections. For each user our detection model derives typical surfing patterns that relate to the topics of interest, frequency of accessing the information, when the information is accessed, etc. Significant deviations from those patterns, particularly when coupled with an event of interest (EOI), such as hit and runs or terrorist attacks, may indicate the subject's active involvement in the event. We provide an outline of the model and the related architecture, which may serve as guidelines for future research.**

Keywords—Security Analitics, Web Intelligence, Forensic Information, User Profiling, Big Data

I. Introduction

Online Internet resources such as portals, news websites, social networks and forums, are an important source of information in the 21st century. While these websites are commonly visited by innocent information consumers, they can be a valuable tool in the hands of outlaws.

Criminals and terrorists use the web to find relevant information before, during and after the criminal/terrorist activity. The common use of the web as a source of information led to the understanding that Google-search-logs, browsing history, emails, and social network activity may provide an effective and relevant evidence for criminal activity [1]. As a result, Computer Forensics has played an increasingly growing part in criminal investigations in recent years, Reports of cases in which the Computer Forensics Analyst has assisted in gathering crucial evidence in homicide trails [2] and even in locating traditional (Non-cyber), criminals [3], are becoming more common. A Research done in University of Virginia [4] even showed the use of Twitter posts for automatic crime prediction.

The use of the Internet by non-cyber criminals has both practical and psychological aspects. Before executing the crime the criminal will likely gather information in order to plan the crime. During the execution of the crime, the criminal will use the Internet to communicate and obtain information about the police investigation through the news coverage of the crime. At the end of the criminal activity, the sensation of "returning to the crime scene" can be solely satisfied by using the Internet. Live and frequent news reports, YouTube videos, with people taking and uploading pictures of the location, provide an all-around coverage of the crime scene thus, satisfying the offender's psychological need, in the safety of his own home.

Legislators around the world understand the value of this information and they are promoting laws that mandate telecommunication data retention and define the legal process in which this data will be available to law enforcement agencies (LEA) [5] [6] [7]. Once all legal requirements are fulfilled, the LEA could inspect and analyze real-time and historical browsing information of suspicious users.

In this paper we propose a way to analyze the information gathered by the Law enforcement agencies. By processing the user's historical surfing information, a LEA can build a user profile representing the user's typical surfing patterns. Significant deviations from those patterns, particularly when coupled with an event of interest (EOI), such as hit and runs or terrorist attacks, may indicate the subject's active involvement in the event.

II. Legal Background

The premise is that the law enables governmental authorities to obtain telecommunication information collected by Internet websites, Internet service providers (ISP), and telephony service providers.

Both the US and the EU have adopted various national laws mandating data retention. Title 18 of the United States Code, Section 2703(f) [8] states that: "*A provider of wire or electronic communications services or a remote computing service, upon the request of a government entity, shall take all necessary steps to preserve records and other records in its possession pending the issuance of a court order or other process*." On March 15th, 2006 the European Union adopted the Data Retention Directive [5]. The Directive requires member states to ensure that communication providers retain necessary data, for a period between 6 months and 2 years.

Besides data retention, which is crucial in posterior forensic analysis, investigators often need real-time interception of Internet traffic. This issue is also covered by legislation. The Communications Assistance for Law Enforcement Act (CALEA) [9] is a United States wiretapping law. CALEA's purpose is to enhance the ability of law enforcement and intelligence agencies to conduct electronic surveillance by requiring that telecommunications carriers and manufacturers of telecommunications equipment modify and design their equipment, facilities, and services to ensure that they have built-in surveillance capabilities, allowing federal agencies to monitor all telephone, broadband internet, and VoIP traffic in real-time.

III. Motivating Scenarios

In this section we will present two motivating scenarios that will help to explain our concept.

A. First scenario: Hit and run accident

Archie is a normative family guy and a respected member of the community. One night, after an urgent discussion with his boss, he drives home. His mind is occupied with the possible consequences of the meeting. He does not notice the pedestrian who storms onto the road out of nowhere. Shocked and confused, Archie leaves the wounded pedestrian on site and drives home (in the USA alone there are over 700,000 hit and run accidents every year [10]).

At home, Archie goes straight to his computer and starts searching the Web for news concerning hit and run accidents in his district. He uses search engines, browses the daily news sites, and also enters Google Street, trying to recall the site of the accident. Archie totally ignores the sports and financial news Websites, which in a usual evening, would be his favorite browsing targets. Instead, he keeps browsing and searching fervently for clues regarding the consequences of his careless driving, trying to figure out his chances of getting caught and his possible defense strategy.

B. Second scenario: Terrorist attack

Alice and Bob are members of the Dark Side Brigades (DSB), a fanatic Sith terrorist group. Bob's mission is to commit a suicide bomb attack at the central subway station of Metropolis. He leaves Alice's place around 9 AM, aiming to reach the central station at 12 AM, the rush hour. Alice, who recruited Bob and coordinated the operation, waits anxiously for confirming evidence of the mission's success. If Bob fails, Alice should resort to plan B, as ordered by her anonymous superiors. She browses the Web fervently, cycling through various daily news Websites in search of breaking news. She also uses search engines with keywords like "subway suicide attack", "Metropolis massacre", and similar strings. As the deadline approaches, Alice's browsing activity accelerates. Then, after finding what she was looking for, Alice closes her computer, prays briefly, and goes on with her daily chores.

IV. Proposed Web Intelligence Architecture

The proposed architecture should take into account legal considerations (described earlier in this paper), which may be quite restrictive in Western democracies. Therefore, appropriate filtering should be applied to preserve the privacy of innocent users in the Internet. The architecture should enable the collection of online data from various Internet Service Providers (ISPs), optionally analyzing the data in real-time, and transmitting the relevant data further to the governmental authority's data store. This concerns huge amounts of data, which should be properly handled both in transport, in storage, and in processing.

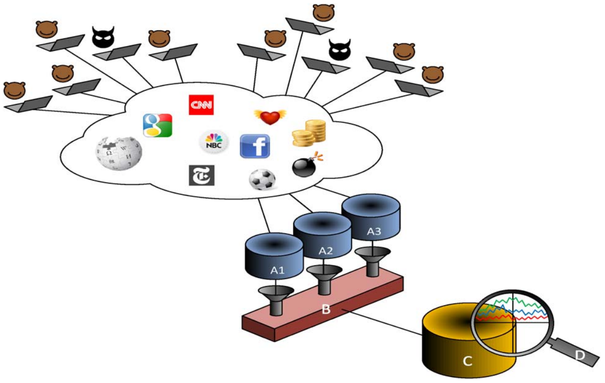

Fig. 1. Proposed Architecture.

As illustrated in Fig. 1, the individual surfers' activities are managed by various ISP’s and are recorded by each ISP (denoted as A1, A2, and A3). The data is filtered according to legal and operational considerations, and transmitted to a relay (denoted as B), which is maintained by the governmental authority. The relay enables real-time analysis where needed. The data is further propagated to a persistent data store (denoted by C), where it can be further analyzed by Big-Data analysis tools (denoted by D).

V. Model

The proposed model, which implements D in the proposed architecture, is based on five parameters that characterize the users surfing behavior: intensity of surfing, frequency of revisiting/refreshing pages, irregular hours of activity, interaction level (passive/active), and diversity of interest topics. These parameters can be derived from the raw browsing data, given a database that contains the complete historic browsing information of the user. The model assumes the existence of a correlation between these activity parameters to a given Event of Interest (EOI).

We suggest that these parameters may indicate the active involvement of the subject in an EOI. Where each parameter in itself may have a limited predictive value, the combination of these parameters may yield an accurate prediction or evidence.

A. Intensity of surfing

This parameter measures the intensity of the user's Internet surfing activities. We measure the browsing intensity value by the number of pages that the user visited in a given time.

When a user shows an increased interest in a given event, we can assume that he will visit related webpages, more intensively than usual.

Surfing intensity may significantly vary between individuals. Consequently, historical data of the user's surfing intensity should be used when searching for anomalies.

B. Frequency of revisiting/refreshing a given page

This parameter measures the number of revisit/refresh operations performed by the user on each page.

Through this information the system may locate stressful behavior, where the user strives for immediate updates regarding his topic of interest. He may repeatedly and frequently revisit the same page, or simply push the 'refresh' button on the browser.

Significant peaks in this parameter may be observed at real-time. However, historical data may improve the predictive value of this parameter by providing its normative range per user.

C. *Irregular hours of activity*

This parameter measures irregular surfing hours and irregular lengths of surfing sessions. Examination of a user's historical data may reveal a regular pattern, concerning his surfing hours. This regularity is governed by overall normal daily routines related to work, leisure and sleep.

This parameter requires analyzing the user's historical data to learn the regular surfing hours and session-lengths. Deviations from such patterns can be found by anomaly-detection methods.

D. *Interaction level (passive/active)*

This parameter measures the level of the user's interaction, ranging from 'low' (passive only), to 'high' (mostly active). In passive surfing the user suffices with reading pages, whereas in active surfing he may chat, write email, commit responses or talkbacks, do Internet shopping, and so on.

Regarding our 'terrorist' scenario, we hypothesize that, as the deadline comes closer, the subject will lower his or her active profile, and will focus on passive consumption of relevant information.

E. *Diversity of interest topics*

This parameter measures the user's range of interest topics, as reflected by the kind of websites he visits.

Surfers are often attracted to diverse topics such as news, sports, music, gaming or finances. Regarding our scenarios, when the subject is focused on an urgent issue, we assume that it will affect his or her surfing pattern, restricting the range of visited sites to a specific topic.

The user's normal diversity measure can be learned from his historical data, using clustering methods. Significant deviations show up as anomalies or outliers.

F. *Correlation with EOI timing*

We assume that our five behavioral parameters are correlated with the timing of the EOI. When the timing of the EOI is known to the investigator, as in forensic investigations, such correlations can provide supportive evidence in a rather straightforward manner.

However, when the timing of the EOI is unknown to the investigator, as in preemptive investigations, the behavioral parameters can still be used for prediction. The False Positive Rate (FPR), in this case may be significant, and should be considered against the risks of missing a true positive, as well as the base-rate of the relevant event.

Regarding the two motivating scenarios, detectable anomalous behavior is expected in the 'hit and run' scenario only after the event, whereas in the 'terrorist attack' scenario it is mainly expected before the event. Consequently, our method is applicable in a posterior forensic investigation in both scenarios, whereas in a preemptive investigation it is not applicable in the 'hit and run' scenario.

VI. CONCLUSION

Big-Data technologies can provide new effective tools for law enforcement authorities against crime and terrorism. The vast amounts of data produced by surfing activity can be stored and monitored, under legal restrictions, and can be analyzed to reveal suspicious behavioral indicators. A person actively involved in a criminal event is assumed to produce anomalous surfing activity before or after the event. Furthermore, this anomalous activity is assumed to produce several detectable behavioral parameters, which are correlated with the timing of the criminal or terrorist event.

In this paper we outlined a conceptual model and architecture for detecting incriminating surfing anomalies. Under proper deployment, our conceptual framework can be used to pinpoint an actively involved criminal or terrorist.

REFERENCES

[1] J. R. Vacca, Computer forensics: computer crime scene investigation, Volume 1, Hingham, Massachusettes: Charles River Media, INC., 2005.

[2] A. Parker, "Search Engine Analysis Vital In Criminal Investigations?," search marketing standard, 9 August 2011. [Online]. Available: http://www.searchmarketingstandard.com/search-engine-analysis-vital-in-criminal-investigations. [Accessed 28 April 2013].

[3] L. Sabin-Wilson, "using the internet to catch traditional (non-cyber) criminals" Threatpost, 10 Oct 2012. [Online]. Available: http://threatpost.com/using-internet-catch-traditional-non-cyber-criminals-101012/. [Accessed 28 April 2013].

[4] X. Wang, M. S. Gerber and D. E. Brown, "Automatic Crime Prediction Using Events Extracted from Twitter Posts," Social Computing, Behavioral - Cultural Modeling and Prediction Lecture Notes in Computer Science, vol. 7227, pp. 231-238, 2012.

[5] "DIRECTIVE 2006/24/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 March 2006," Official Journal of the European Union, vol. 105, pp. 54 - 62, 13 April 2006.

[6] D. McCullagh, "Police: Internet providers must keep user logs," Cnet, 11 July 2011. [Online]. Available: http://news.cnet.com/8301-31921_3-20078653-281/police-internet-providers-must-keep-user-logs/. [Accessed 28 April 2013].

[7] B. Grubb, "Govt wants ISPs to record browsing history," ZDNET, June 11 2010. [Online]. Available: http://www.zdnet.com/govt-wants-isps-to-record-browsing-history-1339303785/. [Accessed 28 April 2013].

[8] "UNITED STATES CODE ANNOTATED TITLE 18. CRIMES AND CRIMINAL PROCEDURE PART I--CRIMES CHAPTER 121--STORED WIRE AND ELECTRONIC COMMUNICATIONS AND TRANSACTIONAL RECORDS ACCESS," Department of justice, [Online]. Available: http://euro.ecom.cmu.edu/program/law/08-732/Crime/StoredCommunicationsAct.pdf. [Accessed 28 April 2013].

[9] "Communications Assistance for Law Enforcement Act of 1994 Pub. L. No. 103-414, 108 Stat. 4279," One Hundred Third Congress of the United States of America, 1994. [Online]. Available: http://askcalea.fbi.gov/docs/calea.pdf. [Accessed 28 April 2013].

[10] "Hit and Run Car Accident," Online lawyer source, [Online]. Available: http://www.onlinelawyersource.com/hit-and-run/car-accident/. [Accessed 28 April 2013].