Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Stress level detection via OSN usage pattern and chronicity analysis: An OSINT threat intelligence module

*Miltiadis Kandias, Dimitris Gritzalis *, Vasilis Stavrou, Kostas Nikoloulis*

Information Security & Critical Infrastructure Protection (INFOSEC) Laboratory, Department of Informatics, Athens University of Economics & Business (AUEB), 76 Patission Ave., Athens GR-10434, Greece

ARTICLE INFO

Article history:
Available online 13 December 2016

Keywords:
Online Social Networks (OSN)
Open Source Intelligence (OSINT)
Privacy
Usage pattern deviation
Stress detection
Insider threat
Threat intelligence

ABSTRACT

Online Social Networks (OSN) are not only a popular communication and entertainment platform but also a means of self-representation. In this paper, we adopt an interdisciplinary approach combining Open Source Intelligence (OSINT) and user-generated content classification techniques with a user-driven stress test as applied to a Greek community of OSN users. The main goal of the paper is to study the chronicity of the stress level users experience, as depicted by OSN user generated content. In order to achieve that, we investigate whether collected data are able to facilitate the process of stress level detection. To this end, we perform unsupervised flat data classification of the user-generated content and formulate two working clusters which classify usage patterns that depict medium-to-low and medium-to-high stress levels respectively. To address the main goal of the paper, we divide user-generated content into chronologically defined sub-periods in order to study potential usage fluctuations over time. To this extent, we follow a process that includes (a) content classification into predefined categories of interest, (b) usage pattern metrics extraction and (c) metrics and clusters utilisation towards usage pattern fluctuation detection both through the prism of users' usual usage pattern and its correlation to the depicted stress level. Such an approach enables detection of time periods when usage pattern deviates from the usual and correlates such deviations to user experienced stress level. Finally, we highlight and comment on the emerging ethical issues regarding the classification of OSN user-generated content.

1. Introduction

Modern mediated means of communication have been radicalised by the introduction of Web 2.0 (O'Reilly, 2009), where users are not mere observers. Instead, they are able to contribute and thus become content generators. As a result, another aspect of users' behaviour and personality is manifested within the context of Online Social Networks. Initial debate on whether or not users present their real self when online has concluded to differentiated results; Amichai-Hamburger and Vinitzky indicated that users tend to transfer their offline behaviour online (Amichai-Hamburger and Vinitzky, 2010) while Tufekci (2013) introduces some conceptual issues which

* *Corresponding author.*
E-mail addresses: kandiasm@aueb.gr (M. Kandias), dgrit@aueb.gr (D. Gritzalis), stavrouv@aueb.gr (V. Stavrou), knikoloulis@gmail.com (K. Nikoloulis).

http://dx.doi.org/10.1016/j.cose.2016.12.003
0167-4048/

question that. These results may be used to further study human behaviour and personality through the prism of OSN usage patterns and what they depict.

These studies have introduced and widened the area of psychosocial characteristics detection and extraction, making it possible to examine users' OSN behaviour (Rogers et al., 2006). This area has been further developed by the use of Open Source Intelligence (Best, 2008) which is defined as "*data produced from publicly available information that is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement*" by the US Dept. of Defense. Further research has indicated that a variety of psychosocial characteristics (e.g., predisposition towards law enforcement, narcissism, various personality traits etc.) may be extracted via Open Source Intelligence (OSINT), produced in the context of OSN and by further data mining (Kandias et al., 2013a) or graph theoretic processing (Kandias et al., 2013b). Regarding possible applications of the results, research indicates that when a person experiences stress, she is more vulnerable to fall prey to third parties, overcome moral inhibitions, or manifest deviant behaviour (FBI, 2012; Greitzer et al., 2012; Shaw et al., 1998). These research papers indicate that stress can be an important indicator when analysing potential insiders, in advanced Threat Intelligence feeds and in forensics analysis of past incidents.

Our major contribution relies on the development of a method that extract results and statistics over the stress level that usage patterns depict through the prism of the "chronicity" of OSN usage. The term chronicity refers to the usage deviations that users may exhibit in the context of OSN over time. To this extent, we selected a Greek community of OSN users (after accepting their informed consent) and classified its members to detect their stress levels over time.

Initially, we conducted unsupervised flat data classification to explore the dataset and formed two working clusters that detect overall usage patterns. The first cluster included mainly users with medium-to-high and high stress level, according to the stress test, while the second cluster included mainly users with low and medium-to-low stress level. In this process, the results of the user-driven stress test were utilised only to calculate the amount of users who were classified in each cluster. One could argue that these results could be linked to the Big-5 trait of neuroticism (Hayes and Joseph, 2003), but analysing this is beyond the scope of this paper and will be addressed in our future work.

In order to study deviations that appeared in the users' OSN usage pattern we utilised machine learning techniques so as to categorise user content into predefined categories of interest (sports, music, politics and miscellaneous). We also examined usage patterns over time in order to form clusters of similar characteristics. For doing so, we developed a series of usage chronicity metrics and fed them to a classification/polling scheme, similar to other relevant classification methods (Veeramachaneni and Arnaldo, 2016), that decides over the usage fluctuations of the OSN. Forming clusters of similar (OSN-usage) time periods enables detection and matching of usage patterns and time fractions. We tested several time-windows, varying from one day to one month, and observed that the time-span of one week suits better the purposes of our research. As a result, rarely observed usage patterns are considered potential deviations from usual, which may mean that the user has experienced real-time events that have consequently affected the OSN usage pattern. We have formed and evaluated eight usage pattern clusters that share similar characteristics and correlated them with users' stress levels.

These methods inherently include a series of ethical issues along with conflicts between stakeholders. Regarding the ethical issues, the methods may interfere with user privacy and lead to a social threat. Regarding the conflicts, it is understandable that law enforcement, crime prevention agencies or even employers or HR managers may request to "connect the dots" and combine information about potential suspects or employees respectively. At the same time, citizens/employees demand that their privacy be respected. Thus, processing user-generated content could be ethically acceptable solely under strict terms and conditions and given the informed consent of the user.

The paper is organised as follows: in section 2 we review the existing literature. In section 3 we describe the dataset collection methodology. In section 4 we describe statistical and content analysis of the dataset. In section 5 we describe the flat data classification method. In section 6 we describe the chronicity analysis and quote our results. In section 7 we discuss the ethical issues. Section 8 concludes the research results.

2. Related work and motivation

The advent of Web 2.0 has contributed to the transformation of the average user from a passive reader into a content contributor. Web 2.0 and OSN have, in particular, become a valuable source of personal data, which are available for crawling and processing, allowing for the extraction of a variety of evaluations through them. Data Science was applied to Psychology issues only recently. The existing literature includes studies that focus on the area of personality, psychopathology and behaviour in relation to OSN usage.

2.1. *Psychology-focused studies*

Regarding personality traits and OSN usage, research highlights that in the context of OSN, users reveal their actual personality rather than self-idealisation (Back et al., 2010). This is further supported by another research that highlights a strong connection between personality and Facebook usage (Amichai-Hamburger and Vinitzky, 2010), supporting that users tend to transfer their offline behaviour online. To this extent, Rosenberg and Egbert (2011) investigate the utility of personality traits and secondary goals as predictors of self-presentation tactics employed by Facebook users. Furthermore, the Big-5 model (Hayes and Joseph, 2003) is studied through the prism of OSN usage. Ross et al. (2009) investigates how the 5-Factor Model of personality relates to Facebook use. They demonstrate a series of correlations between the Big-5 elements and the Facebook usage (e.g. individuals high on the trait of extraversion belong to more Facebook groups but do not have significantly more friends). Similar results are shown by Bachrach et al. (2012) about how users' activity on Facebook relates to their personality, as measured by the standard Big-5 Model, where they attempt to predict personality traits on the basis of Facebook usage