Cybermetrics: User Identification through Network Flow Analysis

Nikolay Melnikov and Jürgen Schönwälder

Computer Science, Jacobs University Bremen, Germany
{n.melnikov,j.schoenwaelder}@jacobs-university.de

Abstract. Recent studies on user identification focused on behavioral aspects of biometric patterns, such as keystroke dynamics or activity cycles in on-line games. The aim of our work is to identify users through the detection and analysis of characteristic network flow patterns. The transformation of concepts from the biometric domain into the network domain leads to the concept of a *cybermetric* pattern — a pattern that identifies a user based on her characteristic Internet activity.

Keywords: Cybermetrics, User Identification, Network Flow Analysis.

1 Introduction

The increasing usage of the Internet in our daily lives led us to believe that Internet citizens have developed a *distinguishable* individual browsing pattern and style. Network flow traces recording personal browsing sessions should contain patterns, representing the users' characteristic cybermetrics. The cybermetric is assumed to reflect user's *priorities* during a browsing activity, the *sequence of the performed steps* at each new Internet browsing session, the *pool of destinations* visited on the Internet and several other features pertaining to that user's characteristic network usage. Having a mechanism for identifying users based on their cybermetrics provides a set of advantages for the purpose of network management, system administration, and security. For example, cybermetrics might be used to grant access to specific services or to verify the identity of a user when she is calling the helpdesk.

In the following section, we state the research questions. We then report some initial experimental results analyzing the impact of the length of network flow traces on the calculation of cybermetrics. We briefly review related work before we conclude our paper.

2 Research Questions

The goal of our research is to identify and distinguish users based on the Internet flows generated by them. The following questions require further investigation:

1. *What is a suitable set of features of a flow trace for user identification?*
2. *What are suitable mathematical methods that can be employed?*

B. Stiller and F. De Turck (Eds.): AIMS 2010, LNCS 6155, pp. 167–170, 2010.

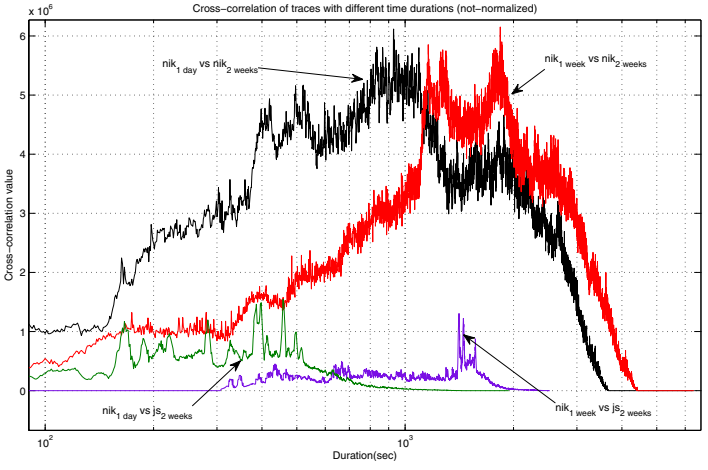

Fig. 1. Cross-correlation of traces for the feature "duration of https connections"

3. *Which thresholds can reliably detect feature similarity (or dissimilarity)?*
4. *What is a scalable approach to automate the user identification process?*

It is evident that a plain comparison of two users' flow traces will result in much noise and would not be a good comparison technique overall. It is therefore necessary to identify a set of features, which have a high potential to differentiate users. The analysis and comparison steps require the usage of proper mathematical methods. While analyzing feature sets, it is important to be able to establish evidence of the similarity of feature sets, or evidence of the dissimilarity of feature sets, or to conclude that no evidence can be derived. Once a suitable user identification technique has been found, we must consider how it can be implemented in a scalable manner.

3 Study of the Impact of the Length of Flow Traces

At the beginning of our research, we wanted to know how cybermetrics may be impacted by the length of flow traces. For our experimental study, we asked several people to collect their personal flow traces by collecting flow records originating from their personal computers.

Considering a flow trace spanning a large number of days, it is expected that the number of longer flows increases compared to shorter traces. Fig. 1 shows the cross-correlation of the feature "duration of https connections" for traces of different lengths and of different users. The plot indicates a strong correlation of this feature for traces coming from the same user. It also indicates a time shift of high correlation values when the lengths of the traces increase. The overall cross-correlation shows a stronger similarity for the traces obtained from the

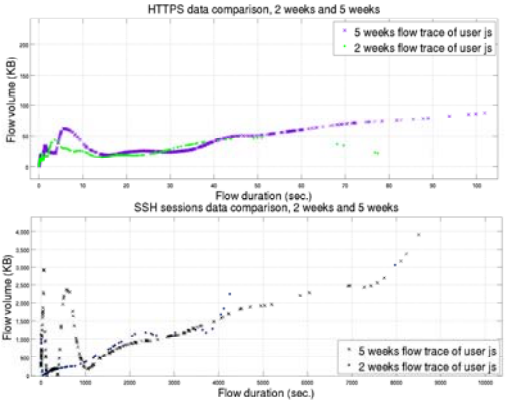

Fig. 2. Scatter-plot of smoothed flow volume (data carried in a flow) vs. flow duration for traces of different lengths for https (top) and ssh (bottom)

same user, `nik`. The cross-correlation of traces from different users (`nik` and `js`) does not indicate strong similarity.

We also wanted to know how the dynamics of the flow volumes depend on the length of the flow traces. The dynamics of two large data sets (coming from the same user) for the two features "volume of https connections" and "volume of ssh connections" is displayed in Fig. 2 (we used the Loess quadratic fit for smoothing the data shown in Fig. 2). In each plot, one curve shows flow data collected over a two-week period (March 10-24, 2009), while the other one shows flow data for five weeks (March 1 - April 4, 2009).

The dynamics of the curves in each plot are quite similar. Essentially for smaller durations (< 5000 seconds), the relationship between the duration and the amount of octets carried is similar for ssh connections. However, a decrease of similarity can be observed for some of the flows that lasted between 500 and 1000 seconds for ssh connections of a five-week long flow trace. The https connections were in general shorter than ssh connections. Furthermore, the five-week flow traces had more occurrences of longer lasting flows. The amount of data carried in flows stayed almost constant independently of the length of the traces. The strong match of the flow volume for most of the flow durations is a good indicator of similarity.

4 Related Work

There are two areas that are closely related to our research. The first area of research deals with user identification methods based on the behavioral and activity features of a user. The idea of user recognition and identification by exploitation of biometric patterns has long been known [1], [2]. More recent studies look at dynamics of certain actions performed by the user — be it an on-line game-play activity [3], which shows that the idle and active times in the

game are representative of the user; a keystroke analysis [4], which provides an impressive 96% correctness rate at user differentiation; or user-mouse interaction dynamics [5], establishing a behavioral characteristic that can be used as an additional security feature.

The second area of research uses passive network traffic monitoring techniques for performance analysis, application type/protocol identification, anomaly and intrusion detections. In [6] the authors propose a novel identification method for revealing Peer-to-Peer traffic. The authors of [7] detect, classify and understand anomaly structures using entropy as a metric of unusual changes in the distribution of traffic features. A more recent study [8] proposes an on-line anomaly detection algorithm that has no prior knowledge about what is normal and abnormal traffic.

5 Conclusion

This paper discusses the possibility of user identification using flow trace analysis. We state our research questions and provide some preliminary results, indicating that the length of the traces can have significant impact on certain flow features.

Acknowledgement

The work reported in this paper is supported by the EC IST-EMANICS Network of Excellence (#26854).

References

1. Holmes, J.P., Wright, L.J., Maxwell, R.L.: A performance evaluation of biometric identification devices. Technical report, Sandia National Laboratories, Albuquerque, NM (1991)
2. Ashbourn, J.: Biometrics: advanced identity verification. Springer, London (2000)
3. Chen, K.-T., Hong, L.-W.: User identification based on game-play activity patterns. In: Proc. of the 6th ACM SIGCOMM Workshop on Network and System Support for Games (NetGames 2007), pp. 7–12. ACM, New York (2007)
4. Bergadano, F., Gunetti, D., Picardi, C.: User authentication through keystroke dynamics. ACM Transactions Information System Security 5(4), 367–397 (2002)
5. Ahmed, A.A.E., Traore, I.: A new biometric technology based on mouse dynamics. IEEE Transactions on Dependable and Secure Computing 4(3), 165–179 (2007)
6. Perényi, M., Dang, T.D., Gefferth, A., Molnár, S.: Identification and analysis of peer-to-peer traffic. JCM 1(7), 36–46 (2006)
7. Lakhina, A., Crovella, M., Diot, C.: Mining anomalies using traffic feature distributions. SIGCOMM Computer Communication Review 35(4), 217–228 (2005)
8. Stoecklin, M.P., Boudec, J.-Y.L., Kind, A.: A two-layered anomaly detection technique based on multi-modal flow behavior models. In: Claypool, M., Uhlig, S. (eds.) PAM 2008. LNCS, vol. 4979, pp. 212–221. Springer, Heidelberg (2008)