Enhancing user awareness during internet browsing*

Bernardo Breve, Loredana Caruccio, Stefano Cirillo, Domenico Desiato,
Vincenzo Deufemia, and Giuseppe Polese

University of Salerno, Via Giovanni Paolo II, 132, Fisciano, Italy, 84084
{bbreve,lcaruccio,scirillo,ddesiato,deufemia,gpolese}@unisa.it

Abstract

With the advent of e-commerce and social networks, people often unconsciously disseminate their sensitive data through different platforms such as Amazon, eBay, Facebook, Twitter, Instagram, and so on. In this scenario, it would be useful to support users with tools providing awareness on how their sensitive data are exchanged. In this paper, we present a visual analytics tool that allows users to understand how their sensitive data are exchanged or shared among different network services. In particular, the proposed tool visualizes the communication flow generated during Web browsing activities, highlighting the providers tracking their data. The tool provides a real-time summary graph showing the information acquired from the network. A user study is presented to highlight how the proposed tool improves the user's perception of privacy issues.

1 Introduction

Nowadays, information is spread over different network channels, and in most cases, users are unaware of how their sensitive data are managed and shared. For example, users have access to many network services, and in order to have complete access to all their features, they must sign an agreement to share their own sensitive information. Moreover, if users grant the consent to process their personal data, they usually have no awareness concerning how and who will manage these data [4]. This could yield different privacy issues, such as the possibility to deal with malicious accounts that might jeopardize users' identities [5].

The application of privacy-preserving policies on the virtual world represents a relevant aspect. This is due to the fact that it is difficult to implement forms of control policies on the Internet, considering its information sharing nature. The privacy preservation problem becomes even more critical when big data need to be processed [10, 15], possibly including web data sources. Several visual languages and visualization techniques have been used to support the management of security policies in the context of Web applications [6, 14].

To define regulations for safeguarding user's personal data, the European community has issued the General Data Protection Regulation (GDPR). The latter is extremely strict, and companies or network providers that do not abide by it are subject to economic fees[1].

In this paper, we propose a tool exploiting visualization techniques to increase users' awareness of how their sensitive data are processed over the Web. In particular, the proposed tool uses an IP network sniffer combined with a NoSQL real-time database [12] for capturing and storing user-related information. Furthermore, the visual interface of the proposed tool relies on several visual metaphors to highlight the packets sniffed during user Web browsing sessions. One of these is an interactive graph resembling a network web of providers, highlighting how they communicate with each other by exchanging user's data. Additionally, the proposed tool

[1] GDPR-https://ec.europa.eu/info/law/law-topic/data-protection

provides several other visual metaphors showing statistics related to mainly involved providers during the user Web browsing session, the network flow in terms of the amount of exchanged packets, and the geo-localization of IP providers on the globe. We performed several user studies to evaluate the capacity of the proposed visual metaphors in terms of increase own user awareness.

The paper is organized as follows. In Section 2 we outline related works. In Section 3 we introduce the architecture of the proposed tool by also describing the underlying technologies. In Section 4 we explain the features of the proposed visual interactive privacy analytic tool, and in Section 5 we discuss the results of user studies. Finally, conclusions and future research directions are proposed in Section 6.

2 Related Work

In this section, we describe several network data visual analytics tools proposed in the literature.

In [1], authors have described 13 network visualization tools, in order to outline their advantages and disadvantages. They employed qualitative coding as part of their research design to extract some metrics from the list of advantages and disadvantages of the tools. Their purpose is to facilitate the analysts during the construction of evaluation methodologies, which they use to measure the effectiveness of visualization tools through usability studies.

In [2], a tool named NetMod is presented. It uses simple analytical models providing designers of large interconnected local area networks with an in-depth analysis of system performances. The tool can be used in environments consisting of thousands of websites. The analytical models and the user interface of NetMod have been tested on a campus-wide network.

MVSec is a visual analytics system supporting analysts to understand how to manage information flows over secure networks [13]. The system permits to perform data fusion activities among multiple heterogeneous datasets, aiming to reduce the effort of the data fusion process, by means of several visual metaphors. The authors defined multiple coordinated views, providing analysts with multiple visual perspectives to characterize loud events, to dig out subtle events, and to investigate relations among events in the datasets. Several case studies have been performed aiming to demonstrate how the system helps analysts in drawing an analytical storyline of networking, and to understand network changes.

Finally, in [8] a prototype 3D visualization system for real-time monitoring of both the status of networked devices (wired, wireless, IoT devices) and the network's dynamics ("pulse") (e.g. configuration, load, traffic, abnormal events, suspicious connections, failed IoT devices, etc.) is presented. Through this prototype, users can visualize the current status of the networks of interest simply and intuitively, and from anywhere on the Internet (even from a mobile device). Furthermore, users can receive alerts through short text or instant messages, whenever something significant occurs on the network.

3 System architecture

In this section, we present the technologies used for building the proposed visual analytics tool. We explain in detail our design choices and how the technologies used are connected. We show how we built the network sniffer, and how this interacts with data streams through ReThinkDB, and how the sniffed packets have been used to compose the proposed visual analytics tool.

3.1 Background

To present the technologies used to construct the proposed tool, in what follows we group them with respect to the component they refer to.

3.1.1 Network sniffer

This component relies on Scapy, which is a Python program that enables users to forge, dissect, emit, or sniff network packets, probes, scans, or network attacks [3]. It provides a DSL (Domain Specific Language) that enables a powerful and fast description of any kind of packet. Scapy allows users to describe a package or a set of packages as layers that overlap each other. The fields of each level have useful default values that can be overloaded. Scapy does not force the user to employ predefined scenarios or templates. Scapy is involved in the receiving phase only, which means that every time a user wants to send packets, a different tool must be involved. When a user probes a network, s/he will send many stimuli, and some of them will be answered. If the user chooses the right stimuli, it is possible to obtain the necessary information from the responses or the lack of responses. Unlike many tools, Scapy will give the user all the information, i.e. all the stimuli s/he sent, and all the responses s/he got. For example, it is possible to probe a TCP port scan, visualize the data in terms of results of a port scan, and then decide whether to visualize the TTL of the response packet. It is not necessary to perform a new analysis every time the user wants to view other data. A common problem in network probing tools is that they try to interpret the answers they got instead of only decoding and giving facts. Saying something like “I received a TCP Reset on port 80” is not subject to interpretation errors. Saying “The port 80 is closed” is an interpretation that can be right most of the times, but wrong in some specific contexts. For instance, some scanners tend to report a filtered TCP port when they receive an ICMP destination unreachable packet. This may be right, but in some cases, it means that the packet was not filtered by the firewall, and there was no host to forward the packet to. We have used Scapy to create an acquisition network script, which is always listening to all possible actions that the user performs during his/her network activities. Our script intercepts all network packets and applies a filter on the communication protocol, selecting only the https and HTTP protocols, and their cookies.

3.1.2 Managing Data Streams

Data streams are managed by means of ReThinkDB, which is a flexible, and scalable database for Web and server development. Like other NoSQL databases, it keeps user data in-sync across client apps through realtime listeners and offers offline support for mobile and Web applications so that the user can build responsive apps working regardless of network latency or Internet connectivity. The ReThinkDB data model supports flexible, hierarchical data structures. The information is stored in JSON documents, organized into collections. Documents can contain complex nested objects in addition to sub-collections. In this type of database, the user can submit queries to retrieve specific documents or all the documents in a collection that matches his/her query parameters. User queries can include multiple chained filters, and combine filtering and sorting. In particular, ReThinkDB uses data synchronization to update data on any connected device. However, it’s also designed to make simple and efficient one-time fetch queries. It offers automatic multi-region data replication, strong consistency guarantee, atomic batch operations, and realtime transaction support. It is a local and/or cloud-hosted NoSQL database, providing access to iOS, Android, and Web apps employing native SDK.

3.1.3 Visualization

In this section, we present the technologies used for implementing the interactive visual interface of the proposed tool. In particular, we implemented a Web application that uses D3.js for building interactive graphics. The latter is a JavaScript library for manipulating data-oriented documents[2]. It achieves visual representations through data loading, data binding, and analytic transformation elements. Unlike graphs generated using Excel, the ones obtainable through D3.js enable developers to define customized mapping rules. Also, physics and force simulations provide a more dynamic representation for the user. Depending on their needs, developers can determine the mapping values to the graphics, such as display color, size, and so on. A D3-based graph is graphically displayed on a Web page by using CSS3, HyperText Markup Language, and Scalable Vector Graphics. It allows dealing with SVG, which is the World Wide Web Consortium specification providing a network vector graphics standard [9]. SVG strictly abides by XML syntax and uses a textual description language to annotate image contents. It is a resolution-independent vector and an image graphic format.

3.2 Architecture

The architecture of the proposed visual analytics tool is shown in Figure 1. In detail, at a low level there are the *Network Sniffer Module* and the *Browsermod Proxy*, which are both responsible for capturing the network traffic. The modules communicate with both the *Data Parser* and the *Database Connector Module*, in order to reorganize and filter the raw data and store them into the *Real-Time Local Database*, which we implemented by using ReThinkDB. In particular, although the latter is only used to store data caught on the web, it offers security mechanisms to prevent unauthorized access. Moreover, web browsing is provided through the *Selenium Driver Module*, which embeds several useful functionalities to manage data retrieved from websites. Furthermore, a Node.js server queries the *Real-Time Local Database* and represents the received data in terms of the proposed metaphors, by using the real-time graphic library (D3.js) presented above. Finally, the structured visual metaphors are presented to the user via the *UI Data Visualization* module.

4 The proposed tool

In this section, we describe our visual analytics tool, named VIPAT, showing its main functionalities. Figure 2 shows a sample of network graph that VIPAT generates to make user aware of the providers that are collecting his/her sensitive data. The graph is constructed in real-time so that the user can analyze the changes occurred when interacting with different providers. In other words, when a user browses Web sites, with different network providers managing his/her personal information, VIPAT captures such information by creating a new node in the graph, as shown in Figure 2. In particular, the azure node within the connected graph represents the user node, the green nodes represent providers using the HTTPS protocol, the red ones represent providers using the HTTP protocol, and the yellow ones represent advertising hosts. Moreover, it is possible to navigate the graph in order to understand how the data are shared by different network providers.

[2] https://d3js.org/

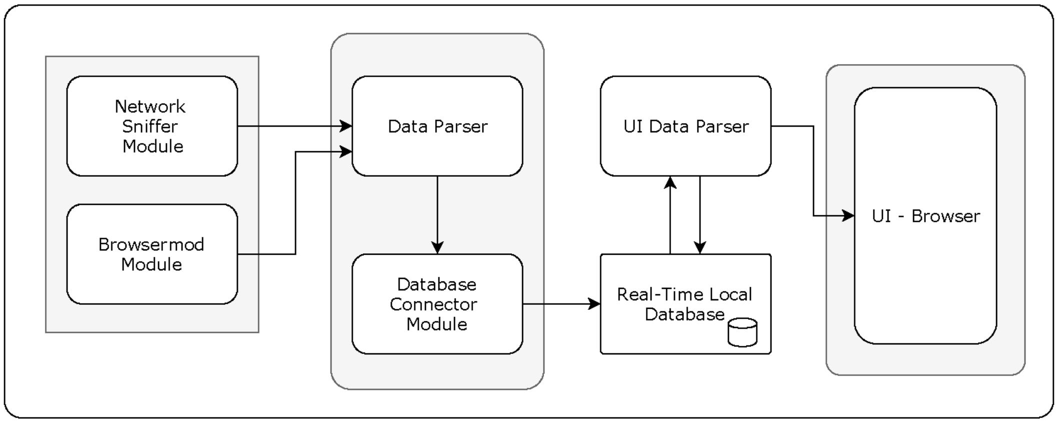

Figure 1: System Architecture.

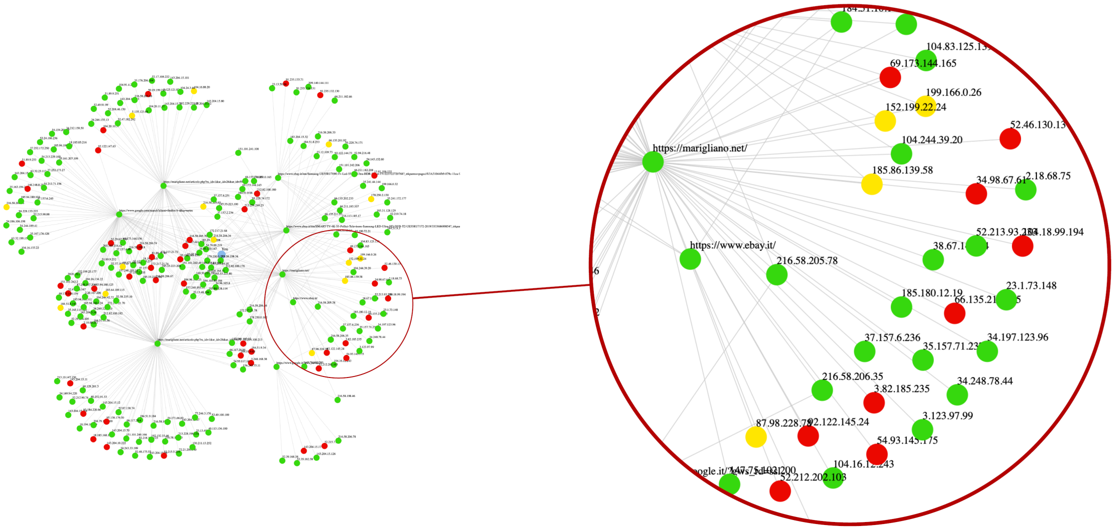

Figure 2: VIPAT network graph representing the information spread on the network during user's browsing sessions.

On the bottom of the visual interface other two interactive charts are provided:

(i) The first one (Figure 3) represents an interactive bar chart showing the top 10 network providers that are taking more advantage from user information, in terms of amount of packets in which they are involved;

(ii) The second one (Figure 4) shows a chart representing the frequency rate at which packets are transferred by service providers on the network each half of a second.

These information allow users to increase their awareness concerning the frequency by which network providers manage their information. The interaction with the charts allows users to retrieve additional information related to the selected network provider. Finally, Figure 5 shows

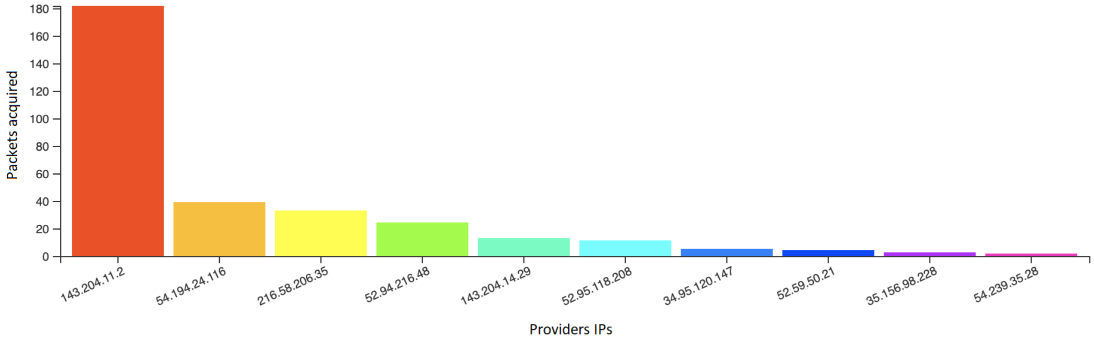

Figure 3: Most contacted providers during user's session browsing.

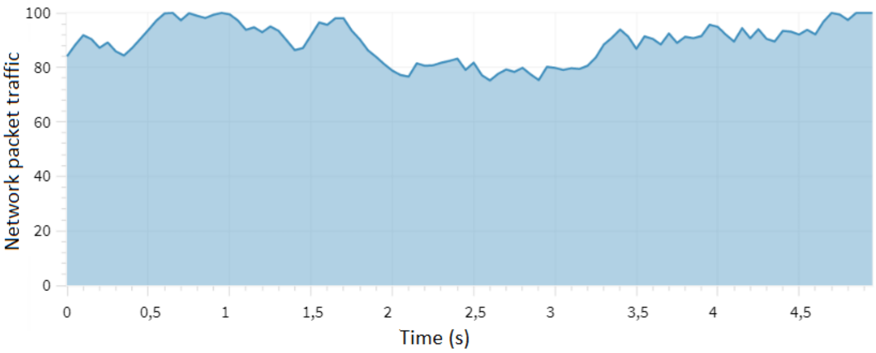

Figure 4: Frequency of the network packets exchanged each 0.5s.

the interface representing a linked globe in which the connections established in the different parts of the world during user Web browsing activities are drawn. In particular, this additional information enables the user to track network communications so that is possible to understand where the network providers are localized.

As it can be seen by the presented interfaces, we decided to adopt simple visual metaphors so that during his/her interaction with VIPAT the user can have a complete overview of the network environment, and be aware of how the different network providers manage his/her information. In fact, one of the most important characteristics of VIPAT is that the user can interact with all of the graphs, by obtaining more specific and/or additional information based on the performed selections.

5 Evaluation

In this section we present the results of a user study whose aim was to evaluate how VIPAT increases the users' awareness about the automatic acquisition of personal data by external providers. The evaluation session has been performed by involving twenty users of different ages in a supervised experiment. Users had different education levels, including high school diploma, master degree, and Ph.D. The user evaluation has been performed in our research

Figure 5: Geolocation of the packet stream worldwide.

laboratory, where users accessed to pre-configured computers having VIPAT installed.

In the first step users filled an initial questionnaire, after which they started performing several Web browsing tasks. In particular, the initial questionnaire contained questions concerning the user's ability to interact with Web applications and other questions related to personal information like age, sex, social occupation, and so on. These questions aimed to asses the following aspects: (1) how much users concern about security and privacy issues, (2) what behavior the users adopt for managing cookies while visiting a web page, and (3) what data generated during the web browsing activities users think they own.

Web browsing tasks have been designed to make a user aware of the spread of his/her own data while performing browsing activities. Five tasks have been defined to be executed by each user, all including actions to perform on Web sites. As an example, a user task included browsing a shopping site and simulating a purchase. Another task required to browse a site of news, and to read the daily news. All users have been supervised during their Web browsing tasks. Moreover, in order to verify the users' interaction with VIPAT, we monitored (*i*) how many times each user opened VIPAT, and (*ii*) how many times s/he interacted with it. After users completed their Web browsing tasks, we asked them to fill out a further questionnaire. The latter included questions concerning user awareness on how his/her personal information is acquired by external providers during Web browsing. In particular, the questions allowed us to evaluate the opinion of the users about whether browsing data are (*i*) safely handled, (*ii*) exploited for commercial purposes, and (*iii*) so frequently exchanged among providers.

For each question, we provided a five-point Likert-type answer format ranging from "strongly disagree" (coded as 1) to "strongly agree" (coded as 5). The initial and the post-task questionnaires share several questions, aiming to monitor whether the users' privacy perception changed after using VIPAT.

Figure 6 shows the users' statistics collected about their concerns on security and privacy issues. It is worth noting that most of the users are quite worried about them and that three students indicated that were not worried although they have knowledge about possible risks. Figure 7 shows the users' statistics about the behavior they adopt with cookies' policies. Most of the users accept them without reading their content. Moreover, no user is willing to leave the website when it is not possible to decline the policy.

Figure 8 shows the statistics about the opinions of the users on the exclusive ownership of the data generated during Web browsing. This question is contained in both questionnaires, thus the figure highlights how the users' opinions changed after using VIPAT. Figure 9 shows the users' answers regarding their opinion about the safe management of navigational data on the Web. Also, this question is contained in both questionnaires, and we can observe that

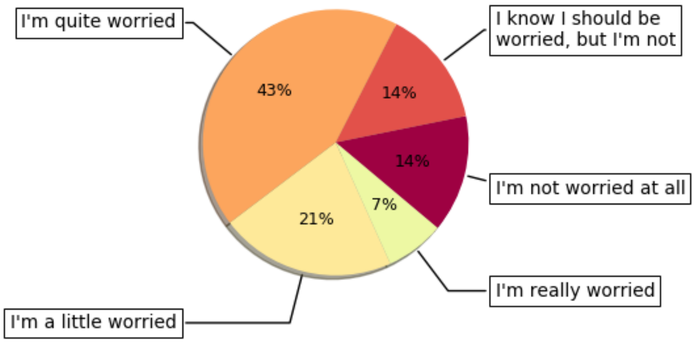

Figure 6: Statistics about the concerns of users on security and privacy issues.

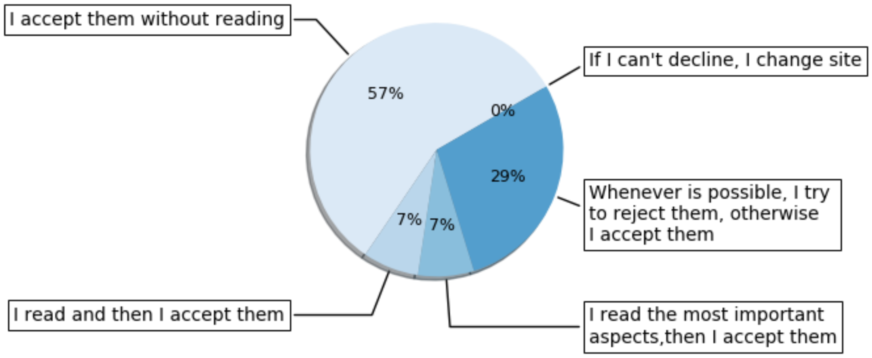

Figure 7: Statistics about the behaviour of the users when dealing with cookies' policies.

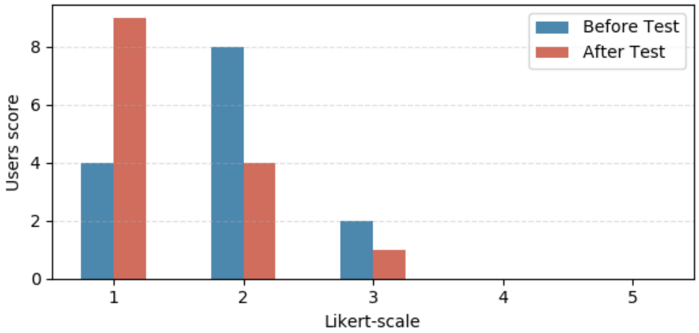

Figure 8: Statistics about users' awareness on the actual ownership of Web browsing data.

VIPAT increased the perception of insecurity about Web data management.

Figure 10 shows the users' answers to the question of how VIPAT changed their data privacy awareness. Most of the users considered VIPAT quite helpful in improving their perception of

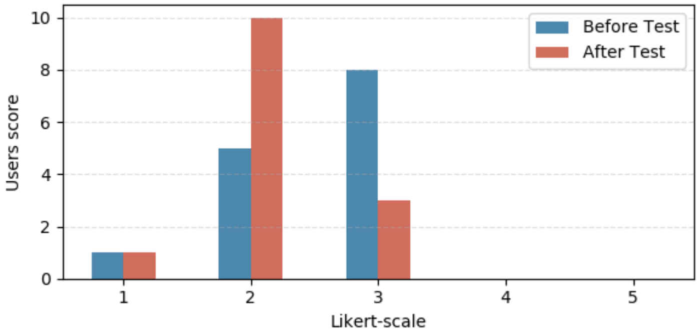

Figure 9: Statistics about users' opinions on how much web browsing data are safely managed.

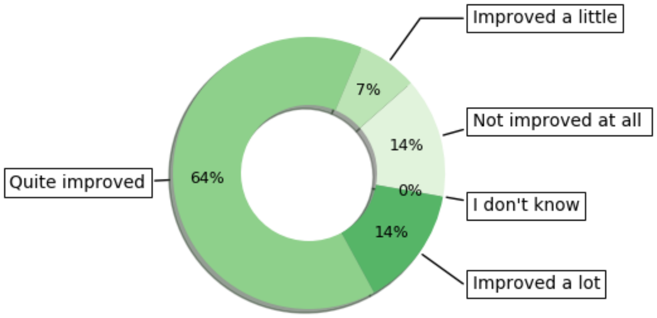

Figure 10: Answers related to the users' awareness on data privacy.

data privacy. In particular, the majority of users, especially those without computer science degrees, gained consciousness on how their data were exchanged without they even know. What they considered really surprising was the fact that during navigation many different providers were communicating with each other, via a third ads-labeled provider, trying to acquiring information on what the user did previously, in order to show a more incisive (and profitable) advertisement.

Figure 11 shows statistics about the users' interactions with VIPAT. The plot highlights how many times users need to get information on the navigation data, looking quickly at or interacting with VIPAT. As can be observed from the results, the number of visualizations is quite different among users. There have been users that very rarely looked at VIPAT (they looked at VIPAT a number of times equal to the number of tasks we assigned them) and others (e.g., User 17) that felt in need to observe the tool a conspicuous amount of times. On average the users interacted with VIPAT 3.6 times during the experimental session, usually at the beginning of the experiment, as the first web page was visited, and after all tasks were completed.

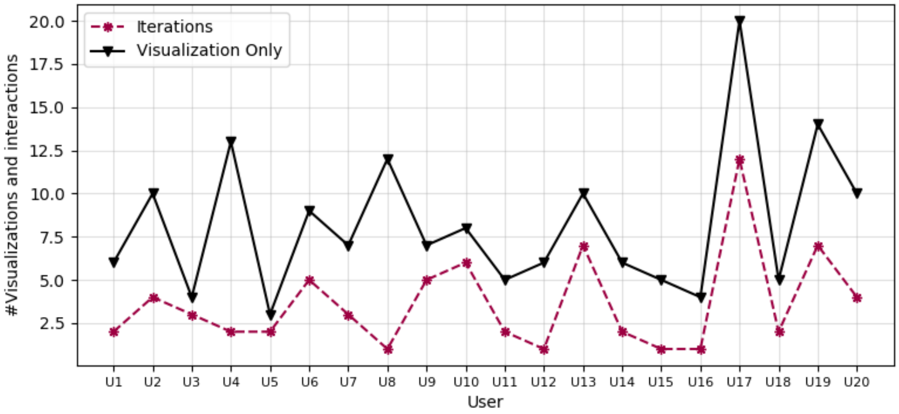

Figure 11: Users' interactions with VIPAT.

6 Conclusion

Nowadays, the Internet is widely used by people of all ages, but not all of them are aware of how their information is shared through different network providers. In this paper, we presented VIPAT, a visual analytics tool that allows users to understand how their information is spread and shared on the network during their Web browsing. In particular, it enables users to observe changes in network environments by means of several interactive graphs, which visualize the providers that share user information together with the frequency rate of exchanged packets.

In the future, we would like to integrate distributed techniques for improving the sniffer of the network traffic module. Moreover, we would like to define additional visual metaphors for adding new information to provide to the user [11]. Finally, we plan to apply the proposed approach to capture the user navigation intents [7].

References

[1] A. E. Attipoe, J. Yan, C. Turner, and D. Richards. Visualization tools for network security. In *Proceedings of Visualization and Data Analysis*, pages 1–8, 2016.

[2] D. W. Bachmann, M. E. Segal, M. M. Srinivasan, and T. J. Teorey. NetMod: A design tool for large-scale heterogeneous campus networks. *IEEE Journal on Selected Areas in Communications*, 9(1):15–24, 1991.

[3] P. Biondi. Scapy documentation (!). https://scapy.readthedocs.io/en/latest/, 2019.

[4] J. Bonneau and S. Preibusch. The privacy jungle: On the market for data protection in social networks. In *Economics of information security and privacy*, pages 121–167. Springer, 2010.

[5] L. Caruccio, D. Desiato, and G. Polese. Fake account identification in social networks. In *Proceedings of IEEE International Conference on Big Data*, pages 5078–5085, 2018.

[6] L. Caruccio, V. Deufemia, C. D'Souza, A. Ginige, and G. Polese. A tool supporting end-user development of access control in web applications. *International Journal of Software Engineering and Knowledge Engineering*, 25(02):307–331, 2015.

[7] L. Caruccio, V. Deufemia, and G. Polese. Understanding user intent on the web through interaction mining. *J. Vis. Lang. Comput.*, 31:230236, 2015.

[8] Z. Constantinescu, M. Vlădoiu, and G. Moise. VizNet Dynamic visualization of networks and internet of things. In *RoEduNet Conf.: Networking in Education & Research*, pages 1–6, 2016.

[9] E. den Heijer and A. Eiben. Using scalable vector graphics to evolve art. *IJART*, 9(1):59–85, 2016.

[10] D. Desiato. A methodology for GDPR compliant data processing. In *Proceedings of the 26th Italian Symposium on Advanced Database Systems*, SEBD '18, 2018.

[11] V. Deufemia, L. Paolino, G. Tortora, A. Traverso, V. Mascardi, M. Ancona, M. Martelli, N. Bianchi, and H. De Lumley. Investigative analysis across documents and drawings: Visual analytics for archaeologists. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '12, pages 539–546. ACM, 2012.

[12] Y. Al Homsi et al. Real-time communication network using firebase cloud IoT platform for ECMO simulation. In *Proceedings of IEEE International Conference on IoT and Green Computing and Comm. and Cyber, Physical and Social Computing and Smart Data*, pages 178–182, 2017.

[13] Y. Zhao et al. MVSec: Multi-perspective and deductive visual analytics on heterogeneous network security data. *Journal of Visualization*, 17(3):181–196, 2014.

[14] M. Giordano, G. Polese, G. Scanniello, and G. Tortora. A system for visual role-based policy modelling. *Journal of Visual Languages & Computing*, 21(1):41–64, 2010.

[15] S. Nicolazzo, A. Nocera, D. Ursino, and L. Virgili. A privacy-preserving approach to prevent feature disclosure in an IoT scenario. *Future Generation Computer Systems*, 105:502–519, 2020.