How old users are? Community analysis.

Valerii Oliseenko[a,b], Anastasia Korepanova[a,b]

[a] *St. Petersburg State University, Mathematics and Mechanics Faculty, Universitetskaya Emb., 7-9, St. Petersburg, 199034, Russian Federation*
[b] *St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 14-th Linia, VI, No. 39, St. Petersburg, 199178, Russian Federation*

Abstract

The presented work is devoted to the development of an approach to the age inference for users on the online social network VK.com. This approach is based on an analysis of the communities that the user is a member of. The proposed method is based on the hypothesis that users join some of those communities that contain data about class, graduation from school, data about the group at the university, etc. and those that have users' peers as members. The proposed approach allows for the age inference. The obtained results can benefit research on online social networks, as well as the task of discovering a user vulnerability through social networks.

Keywords 1

User age inference, information security, social networks, fuzzy algorithms, social engineering attacks.

1. Introduction

Experts note the steady annual increase in the number of cybercrimes on information systems. [1]. Most of the attackers carry out their attacks using social engineering methods [2], despite this in research more attention is paid to the technical and software side of the issue [3, 4, 5, 6, 7]. Thus, there is a strong need on the researches focused on the protection of users from attacks that abuse their vulnerabilities based on personal characteristics. The company Retruster points to the rapidly increasing threat of social engineering attacks, for example, the latest report showed, that the number of phishing attacks alone increased by 65% in 2019 [8].

In order to develop approaches to improve the security of an information system from social engineering attacks, the vulnerability of its users to such attacks should be assessed, in [9, 10] authors present the developments for this assessment. Presented approaches base on the users' vulnerability profiles that reflect their vulnerabilities to different acts of the attack [10]. User accounts in online social networks can be used as the source of data for its construction, as they contain personal data, a list of the user's friends, information posts, subscriptions, etc. Unfortunately, in many cases, the profile information in user accounts may be deliberately hidden or not specified. For example, in the social network VK.com user can hide or not specify a lot of information: list of friends, city of residence, date of birth, education, career, etc.

Missing personal data may cause problems in the analysis of the user's accounts in the tasks associated with determining vulnerabilities of the user or searching for the user's accounts in different online social networks [11, 12, 13, 14]. In this regard, the task of missing personal data inference become relevant. The focus of this paper is on inferring the age of the user by analyzing the communities that the user is a member of. VK.com was chosen as an online social network for testing the effectiveness of proposed methods, as according to Brand Analytics estimates, [15], it is the most popular online social network in Russia with 38.2 million users. The theoretical significance of the work

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia
EMAIL: vdo@dscs.pro; aak@dscs.pro

lies in the created method for user age inference. The practical significance lies in the possible application of the proposed method to different researches on online social networks and in the field of automating the assessment of a user’s protection against social engineering attacks.

2. Literature review

The problem of the age inference of an online social network user is not new. Authors of the work [16] present some systematization and analysis of approaches to the personal data inference. The authors of this paper offer the following classification of approaches to inferring a user's personal data:

- Methods based on the analysis of the user’s behavior in an online social network;
- Methods based on the analysis of the user's social circle, namely friends and their interactions;
- Methods based on the analysis of profile data from several online social networks.

Methods from the first group include status update activity analysis [17], likes analysis [18], analysis of published images [19] and others. The presented methods and approaches are also applicable to the user's age inference.

Methods from the second group usually use different social graph models [20, 21]. Methods based on probabilistic models of Bayesian networks are also can be used [22]. One of the easiest methods to infer a user’s age is to choose a mode of the user’s friends’ age distribution. This method is presented in [23].

Methods based on the analysis of profile data from several online social networks allow inferring age by finding the user's account on other online social networks and checking whether the user has provided his or her age there. There are some methods applicable to determining the accounts of one user in different online social networks. Thus, for example, authors of [24] present the assumption of similarity in the style of the user's text in different networks, in [25, 26, 27] the similarity of the user's publishing activity in different online social networks is discussed, and in [28] the proximity of social circles represented as friend lists, etc. is considered.

This work is a part of a general project aimed to improve information systems’ security by automating user vulnerability to social engineering attacks assessments. These assessments are based on users’ personality features that can be inferred from the data extracted from users’ online social networks accounts. [10].

3. Problem statement

The proposed article solves the problem of user age inference by analyzing communities that the user is a member of. We propose an approach based on the assumption that users usually subscribe to some groups to communicate with their peers. Such groups in their names can contain data about graduation from school, data about group at the university, etc.

Accounts of users from the online social network VK.com serve as an input to this task. The result of solving the task is the inferred age of the users.

4. Methods

1) Selection of communities

The method for determining presumptive peer communities relies on the semantic analysis of community names. We choose those communities from the user's subscriptions, which names contain the following words or their forms: "class", "group", "graduation", and/or Russian designation of classes: "6б", "5 -а ", etc. In order to reduce the number communities that do not belong to educational institutions, but using this method can be falsely presumed to belong, we define the words that the name of a community should not contain: "master class", "musical ", etc. There is also a limit on the number of community members: no more than 100.

2) Community analysis

- The user's age inference by drawing the date from the name of the community. We look for the word "graduation" or designation of the class in the name of the community and the year. For example, "graduation 2014", "5-A 2018". Based on this data, the estimated age of the user is calculated.

- Age inference through the analysis of the ages of community members. Accordingly, to the main assumption that lies behind the community selection method, we identified peer communities in the first step, thus the majority of each community members should have similar age. The second step is to determine the age that is the most common among the subscribers, calculate the total number of subscribers of this age, as well as two years older or younger. The amount is divided by the total number of community members with an open age (date of birth). If the resulting ratio is greater than a certain threshold *i*, then the age of the user is considered to be the age that the majority of the community members have.

5. Experiment Result

We collected a dataset to test the proposed methods on real data and assess their effectiveness. This dataset is composed of 13,500 VK.com accounts. Every account has open user age info. We used the following metric to evaluate quality and effectiveness of the proposed methods: Accuracy = $\frac{T}{N}$, where T denotes the number of accounts for which the age is correctly determined, N denotes the total number of accounts for which the method was applied.

Firstly, we applied the proposed method for the selection of communities, thus identifying the presumptive peer communities. The total of accounts that has subscriptions to communities that meet the requirements turned out to be 2525, i.e. their share was about 0.18 of the original datasets.

The selected communities were analyzed using two proposed methods. The application of the first method showed the following results: there were 214 users subscribed to communities containing dates in their names that were available for the analysis by the first method, i.e. about 0.08 from the users selected in the previous step and about 0.015 from the initial set of accounts. The accuracy of the first method on the dataset is 0.53.

The results of applying the second method are as follows: the number of users for whom the second method can be applied, namely those that are members of communities with suitable age distribution, and the accuracy of its application depends on the parameter *i* (Figures 1, 2, 3).

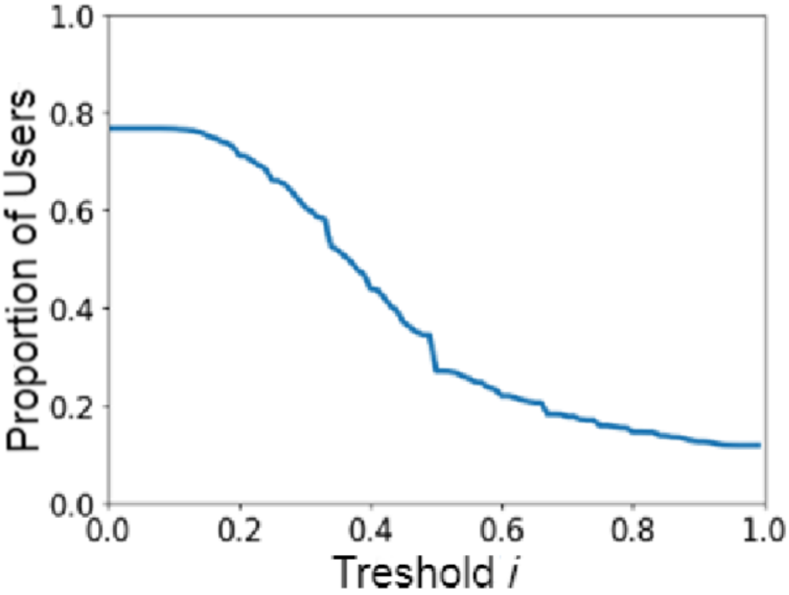

Figure 1: The proportion of users to whom the method can be applied.

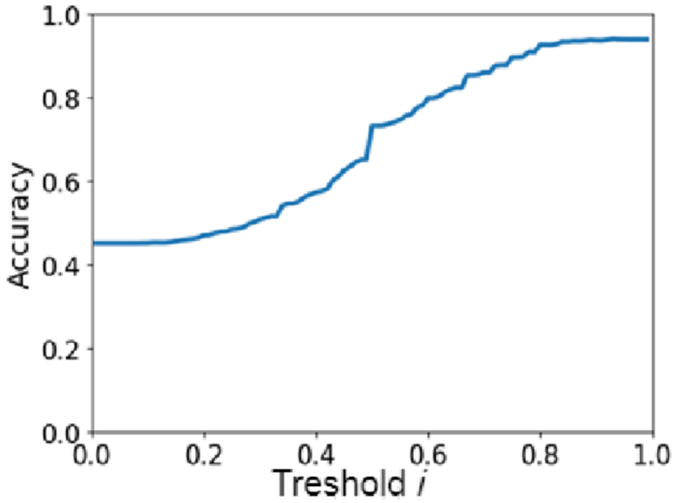

Figure 2: The accuracy of the age inference method. Matching up to three years

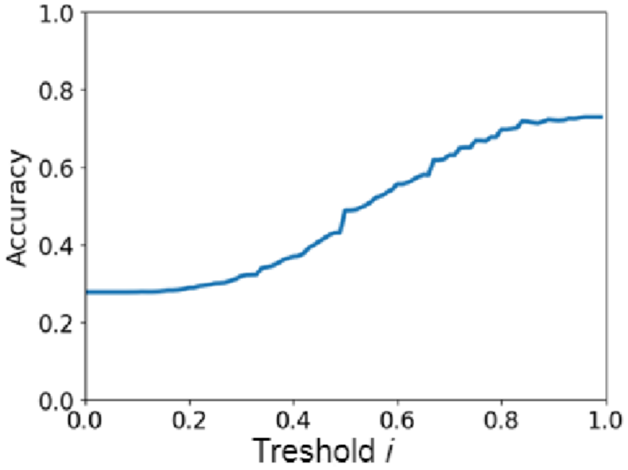

Figure 3: The accuracy of the age inference method. Exact match

The highest value of the threshold *i* correspond to the highest accuracy of the proposed method on the test dataset. However, the higher the value of *i*, the smaller the proportion of users to whom this method is applicable. If *i* is greater than 0.6, only 0.2 of the users selected at the first step, i.e. 0.036 of the total collected datasets were members of communities suitable for analysis. However, the proposed method can be used in conjunction with other methods. In this regard, we plan to continue this research on approaches to the user's age inference, and construct a model combining several methods to increase their joint effectiveness.

6. Conclusion

This paper presents an approach to the age inference of VK.com users. Two methods of inference of the user's age were presented. The fist utilizes semantic analysis of community names, the second focuses on the analysis of the distribution of ages of community members. We have tested the presented methods on a real dataset obtained from an online social network. This study can benefit other research on online social networks and improve other methods and approaches for the age inference of an online social network user. During the further studies, we plan to test the possibility of constructing a model that will combine the proposed methods of age inference with other methods to increase the total effectiveness. We consider the development of methods for finding the optimal value of the parameter *i* as well.

7. Acknowledgements

The research was carried out in the framework of the project on state assignment SPIIRAS № 0073-2019-0003, with the financial support of the RFBR (project No. 18-01-00626, No. 20-07-00839).

8. References

[1] Verizon Data Breach Investigations Report 2018, 2019. URL: https://www.researchgate.net/profile/Suzanne_Widup/publication/324455350_2018_Verizon_Data_Breach_Investigations_Report/links/5ace9f0b0f7e9b18965a5fe5/2018-Verizon-Data-Breach-Investigations-Report.pdf?origin=publication_detail

[2] Ptsecurity — Current Cyber Threats: Results for 2019, 2020. URL: https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2019/

[3] V. V. Borisov, A. S. Fedulov, Model of intellectual risk management while ensuring information security of processes in complex social and technical systems, Neurocomputers and their applications: thesis of XVII All-Russian scientific conference, Moscow, 2019.

[4] A. A. Branitskiy, I. V. Kotenko, Analysis and Classification of Methods for Network Attack Detection. SPIIRAS Proceedings, 2, 45 (2016) 207-244. doi:10.15622/sp.45.13.

[5] I. Kotenko, I. Parashchuk, Verification of unreliable parameters of the malicious information detection model, Vestnik of Astrakhan State Technical University. Series: Management, computer science and informatics, 2 (2019) 7-18. doi:10.24143/2072-9502-2019-2-7-18.

[6] M. Conti, L. V. Mancini, R. Spolaor, N. V. Verde, ”Analyzing Android Encrypted Network Traffic to Identify User Actions,” in IEEE Transactions on Information Forensics and Security, 11, 1 (2016) 114-125. doi: 10.1109/TIFS.2015.2478741.

[7] Y. Zhang, L. Y. Zhang, J. Zhou, L. Liu, F. Chen, X. He, ”A Review of Compressive Sensing in Information Security Field,” in IEEE Access, 4 (2016) 2507-2519. doi: 10.1109/ACCESS.2016.2569421.

[8] Phishing Statistics and Email Fraud Statistics, 2019. URL: https://retruster.com/blog/2019-phishing-and-email-fraud-statistics.html

[9] A. Khlobystova, M. Abramov, A. Tulupyev, An Approach to Estimating of Criticality of Social Engineering Attacks Traces. In: Dolinina O., Brovko A., Pechenkin V., Lvov A., Zhmud V., Kreinovich V. (eds) Recent Research in Control Engineering and Decision Making, ICIT 2019, Studies in Systems, Decision and Control, 199 (2019). Springer, Cham. doi:10.1007/978-3-030-12072-6_36.

[10] M. V. Abramov, T. V. Tulupyeva, A. L. Tulupyev, Social engineering attacks: social networks and user security estimates. SPb: PC GUAP, 2018, ISBN 978-5-8088-1377-5.

[11] A. A. Korepanova, V. D. Oliseenko, M. V. Abramov, A. L. Tulupyev, Application of Machine Learning Methods in the Task of Identifying User Accounts in Two Social Networks, Computer tools in education, 3 (2019) 29-43. doi:10.32603/2071-2340-2019-3-29-43.

[12] J. Henriksen-Bulmer, S. Jeary, Re-identification attacks—A systematic literature review, International Journal of Information Management, 36, 6 (2016) 1184-1192.

[13] Y. Li, Z. Su, J. Yang, C. Gao, Exploiting similarities of user friendship networks across social networks for user identification, Information Sciences, 506 (2020) 78-98.

[14] A. Esfandyari, M. Zignani, S. Gaito, G. P. Rossi, User identification across online social networks in practice: Pitfalls and solutions. Journal of Information Science, 44, 3, 377-391. doi:10.1177/0165551516673480.

[15] Social networks in Russia: research of Brand Analytics, 2019 URL: https://popsters.ru/blog/post/auditoriya-socsetey-v-rossii.

[16] T. V. Sokolova, A. M. Chepovskiy, Problem of inferring profiles of social networks users, Voprosy kiberbezopasnosti, 4 (2019) 88-93.

[17] H. A. Schwartz, J. C. Eichstaedt, M. Kern, L. Dziurzynski, S. M. Ramones, M. Adrawal, A. Shah, Personality, gender, and age in the language of social media: The open-vocabulary approach, PloS one, 8, 9(2013).

[18] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proceedings of the National Academy of Sciences, 110, 15 (2013).

[19] Q. You, S. Bhartia, T. Sun, J. Luo, The eyes of the beholder: Gender prediction using images posted in online social networks, 2014 IEEE International Conference on Data Mining Workshop.

[20] A. G. Gomzin, S.D. Kuznetsov, A method of automatically estimating user age using social connections. Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS), 6 (2016),171-184. doi:10.15514/ISPRAS-2016-28(6)-12.

[21] V. O. Chesnokov, Predicting the attributes of a social network user profile by analyzing the communities of the graph of his immediate environment, Vestnik MGTU im. N.E. Bauman. Instrument making series, 2 (2017). doi: 10.18698/0236-3933-2017-2-66-76.

[22] N. A. Kharitonov, A. G. Maximov, A. L. Tulupyev, Algebraic Bayesian Networks: Naïve Frequentist Approach to Local Machine Learning Based on Imperfect Information from Social Media and Expert Estimates. In: Kuznetsov S., Panov A. (eds) Artificial Intelligence, RCAI 2019, Communications in Computer and Information Science, 1093 (2019), Springer, Cham. doi:10.1007/978-3-030-30763-9_20.

[23] N. E. Slezkin, M. V. Abramov, T. V. Tulupyeva, Approach to reconciliation of information system user's meta-profile based on data from social networks websites, Proceedings of scientific papers Fuzzy Technologies in the Industry, FTI-2017. Ulyanovsk, UlSTU, 2017, 394-399. ISBN 978-5-9795-172

[24] R. Zafarani, H. Liu, Connecting users across social media sites: a behavioral-modeling approach, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013.

[25] O. Goga, H. Lei, S. H. Krishnan, G. Friedland, R. Sommer, R. Teixeira, Exploiting innocuous activity for correlating users across sites, Proceedings of the 22nd international conference on World Wide Web, 2013.

[26] L. Nie, L. Zhang, M. Wang, R. Hong, A. Farseev, T. Chua, Learning user attributes via mobile social multimedia analytics ACM Transactions on Intelligent Systems and Technology, 8 (2017), art. no. 36. doi: 10.1145/2963105

[27] N. Z. Gong, B. Liu, Attribute inference attacks in online social networks, ACM Transactions on Privacy and Security, 21 (2018), art. no. 3. doi: 10.1145/3154793

[28] N. Korula, S. Lattanzi, An efficient reconciliation algorithm for social networks, Proceedings of the VLDB Endowment, 7, 5 (2014).