Approaches To Merging Linguistic Values — Users Relationships

Anastasiia O. Khlobystova[a,b], Alexander L. Tulupyev[a,b]

[a]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
[b]St. Petersburg State University, St. Petersburg, Russia

Abstract

Social engineering attacks based on the human factor have long been the most frequently used in violation of the information security policies. One of the ways to increase the organization's level of protection against social engineering attacks is building a social graph of the organization's employees and its analysis. The nodes of such graph associated with users of the information system, and edge designate the relationships between them. Moreover, this kind of information can be obtained by analyzing social networks. However, often users have accounts in different social networks, and the information presented in them is different. The purpose of this article became to propose approaches to merging probabilistic estimates of the relationship between users, which are linguistic values of linguistic variable "type of relationship". The theoretical significance of the results lies in the proposal of new approaches to the merging of probabilistic estimates of linguistic variables, the practical significance consist in creating the basis for further analysis of the social graph of the organization's employees, in particular, for detecting the most critical trajectories of attack development or solving backtracking tasks of social engineering attacks, e.i. the investigation of cyber crime committed by using social engineering techniques.

Keywords

social engineering attacks, interaction intensity estimates, linguistic variable values, soft computing, merging social networks

1. Introduction

For a long time in the field of information security, one of the least developed sections remains the issue of user protection from cyberattack [1]. Over the years, information security specialists have been developing technical means of protection against hack, for example, against DoS attacks [2], packet sniffing [3], special programs (viruses, worms, trojans) [4] etc. At the same time, organizations are still being unprotected due to the lack of employees' awareness about cybersecurity and, in particular, about social engineering attacks. We consider social engineering attack as a set of applied psychological and analytical methods which malefactors use for users' motivation in terms of public or corporate network in relation to violations of the settled rules and politics in the field of information security [5].

Russian Advances in Artificial Intelligence: selected contributions to the Russian Conference on Artificial intelligence (RCAI 2020), October 10-16, 2020, Moscow, Russia

aok@dscs.pro (A.O. Khlobystova); alt@dscs.pro (A.L. Tulupyev)

This information is confirmed by the statistics of cybersecurity incidents both in Russia [6] and abroad [7]. Moreover, even a small group of malefactors can commit more than 2 thousand cybercrimes in a short period of time, such information is proved in source [8], which describes about the detention of three unemployed young people from St. Petersburg who stole more than 4 million rubles within six months, by creating phishing "clones" of famous brands. Also Ria News have reported the number of cybercrimes in Russia over 5 years has grown by more than 25 times [9], and one of the most common types of cybercrime in Russia are social engineering [10]. In addition, one of the largest banks in Russia expects that in 2021 the Russian economy will lose up to 7 trillion rubles from cybercrime [11].

This requires the development of effective and robust methods, models, methodologies and automated tools against hack with applied social engineering. And since social engineering attacks are directed at people, the overall research direction consists in increasing the level of protection of users of information systems from social engineering attacks.

1.1. Prerequisites for research

One of the important steps to achieve this goal is to analyze the security of users from such threats. Automation methods for building estimates of the security of users of information systems from social engineering attacks were proposed in [5, 12]. In particular, the authors developed a set of models "critical documents — information system — user — malefactor" used to analyze user security, indirectly, critical documents and simulate scenarios of the social engineering attack.

One of the most used sources of information in social engineering attacks is social networks, so according to [13] attacks on accounts on social networks are considered very effective. However, malefactor is not confined to an account on only one social networks. Often users use different social networks for different purposes. In this regard, for the analysis of user security, it is important to consider information from different social networks. This in turn raised questions about aggregation and merging of data. The tasks of merging user accounts from different social networks were discussed in [14, 15, 16]. But beyond merging user profiles, it is also necessary to compare estimates of the probability of attack propagation for social graphs built on different social networks [17]. That is, in other words, it is necessary to compare the estimates of the probability of attack propagation. Thus, the purpose of this study is to propose approaches to the partial merging of social graphs, that is to compare estimates of the probability of attack propagation for different social graphs obtained on the basis of data from different social networks.

1.2. Related Work

The "Friend-of-a-Friend" technology, merging social graphs from various social networks into one database, is described in [18]. However, the authors do not consider the problem of merging inconsistent data that belong to the same category. The analysis of social graphs in the context of rumour spreading in social networks is also considered in [19]. Its authors propose an approach to identifying and blocking nodes that are most likely to disseminate a large amount of false information. The work may be useful in developing approaches to the analysis

of a social graph in the context of identifying the most probability paths for the spread of the social engineering attack, but in it only the values "friendship", "follows" and "subscriptions" are considered as edges of a social graph, which complicates the applicability to the present study. The purpose of the study [20] is to propose a scheme for matching user content identifiers that are publicly available on social networks. Namely, the authors propose a method based on natural language processing and text mining. The study is relevant and useful when aggregating data from social networks, but it does not address the issues of combining information about the interaction between users. A basis for this research was the work of [5, 21] in which approaches to the construction and analysis of the social graph of employees were proposed, methods for quantifying linguistic variables (the relationships between users associated with edges of social graph employees of the organization).

2. Problem statement

Let the user's profile U_i and list of friends be known in the first social network $F = \left\{ U_{j_1}, ..., U_{j_n} \right\}$ $(j_k \neq i, 1 \leq k \leq n)$, also in another social network, this user corresponds to profile U_i' and list of friends $F' = \left\{ U'_{j_1}, ..., U'_{j_n} \right\}$ $(j_k \neq i, 1 \leq k \leq n)$. In this case, the user profile $U_{j_k} \in F$ $(j_k \neq i, 1 \leq k \leq n)$ corresponds to the user profile $U'_{j_k} \in F'$ $(j_k \neq i, 1 \leq k \leq n)$. Obtaining this information relates to the task of merging user profiles in different social networks and has successful solutions [14, 22].

It is known that E_{ij_k} corresponds to relationship between users U_i and U_{j_k} $(j_k \neq i, 1 \leq k \leq n)$, E'_{ij_k} relationship between U_i' and U'_{j_k} $(j_k \neq i, 1 \leq k \leq n)$. In this case E_{ij_k} and E'_{ij_k} can be different from each other. An example of this is social graphs on the "VKontakte" and "Instagram": in "VKontakte" users may be relatives (they are indicated at each other in the corresponding public lists of friends), and in "Instagram" one of them may be follows for other. Figure 1 illustrates an example of merging user profiles and their relationships in two different social networks. At the moment, we do not consider the comparison of relations between the friends of the user, since for them everything will be the same.

According to [5] when constructing a social graph of the organization's employees for the purpose of subsequent analysis to identify the most vulnerable places to social engineering attacks, each relationship between users (E_{ij_k}) is associated with a probabilistic estimate (p_{ij_k}), obtained by analyzing user interactions in social networks (assignment to any category from the list of friends, as well as the availability of shared photos, information about likes, reposts, comments, etc.). Based on the study [21], p_{ij_k} can be obtained by quantifying the types of user relationships. For example, on a social network, information about the relationship between users can be obtained by looking at the public list of friends of the user or the "personal information" section on the profile's main page. Similar actions can be performed with other social networks.

So with this approach, E_{ij_k} is a linguistic variable "type of relationship", which characterize 1-to-1 relation between two users of this network.

Let us remember that by a linguistic variable is meant a variable whose values are words or sentences in a natural or artificial language [23, 24]. Linguistic variable is a quintuple $(L, T(L), U, G, M)$ in which

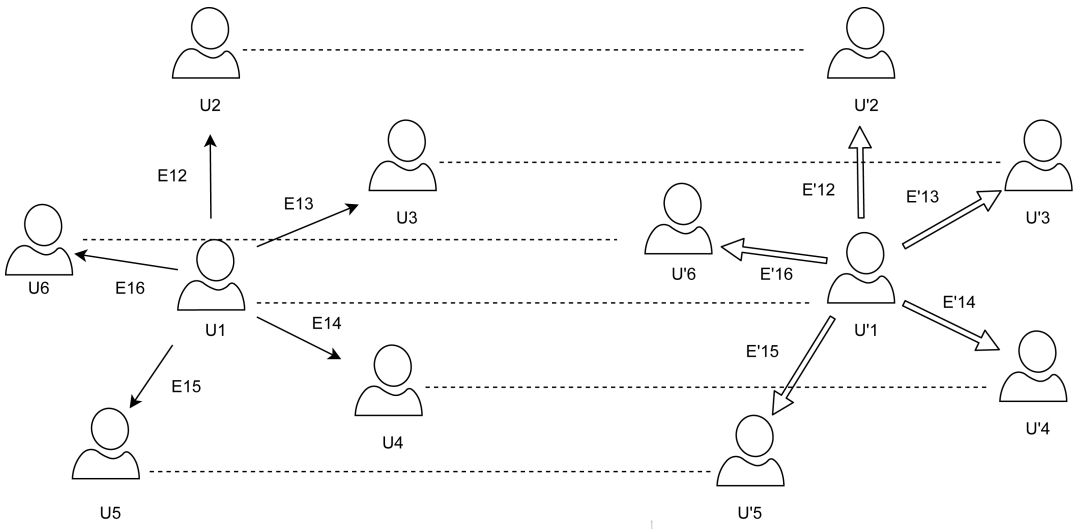

Figure 1: Merging example of communication between users.

- L is the name of the variable;
- $T(L)$ is the term-set of L, that is, the collection of its linguistic values;
- U is a universe of discourse;
- G is a syntactic rule which generates the terms in $T(L)$;
- M is a semantic rule which associates with each linguistic value X its meaning, $M(X)$, where $M(X)$ denotes a fuzzy subset of U.

Thus, for the case under consideration:

- L is “type of relationship”;
- $T(L)$ depends on the social network in question, for example, if the social network is “VKontakte” in our study $T(L)$ = “Friends”+ “Best friends”+ “Colleagues”+ “School friends”+ “University friends”+“Family“+ “Grandparents“+ “Parents“+ “Siblings“+ Children“+ “Grandchildren“+ “In a relationship“+ “Engaged“+ “Married“+ “In a civil union“+ “In love“+ “It’s complicated“;
- U is set of values from $[0, 1]$, denote the strength of the relationship between users;
- G is determined depending on the social network in question;
- M is a modified method by Khovanov described in [22, 21].

Thus, the purpose of this article became to propose approaches to merging probabilistic estimates of the relationship between users, which are linguistic values of linguistic variable "type of relationship".

3. Approaches to merging probabilistic estimates of the relationship between users

This section provides approaches to merging probabilistic estimates of user relationships.

3.1. Highlighting the strongest communication

This approach is based on the assumption that users can subconsciously select one of the social networks and be more active in it, including posting more detailed information about themselves and their social environment. In addition, often different social networks are used for different purposes, in this regard, in one of them, for example, there will be more information related to the work of the user, and in the other with family ties.

Based on the approach to quantification of numerical estimates proposed in [22, 21], we will consider the numerical probabilistic estimates characterizing the relationship between users to be known (Table 1 and Table 1).

Table 1
Probabilistic estimates of the the linguistic values variable “type of relationship" from social network "VKontakte"

Types of relationships	Mapped estimate of probability
Friends	0.2938
Best friends	0.7838
Colleagues	0.4074
School friends	0.4443
University friends	0.3686
Family	0.3641
Grandparents	0.2507
Parents	0.3421
Siblings	0.4398
Children	0.41
Grandchildren	0.3474
In a relationship	0.3075
Engaged	0.3107
Married	0.3793
In a civil union	0.3223
In love	0.4189
It’s complicated	0.1922

In this case, we can compare the linguistic values of the relationship variable, revealing a connection with a larger rating and designating it as a stronger one. Then match this connection to the social graph edge. For example, let us want to compare the relationship between the profiles of “VKontakte” and “Instagram” users. It is noted that “VKontakte” users are connected by relationships E_{ij} — “school friends” (probabilistic assessment corresponds to this type of relationship is $p_{ij} = 0.4443$), in “Instagram” E'_{ij} marked as “i like j” ($p'_{ij} = 0.34$). Then the “school

Table 2
Probabilistic estimates of the the linguistic values variable “type of relationship" from social media "Instagram"

Types of relationships	Mapped estimate of probability
I follow	0.48
I like	0.34
I comment	0.4
I have photo with tag X	0.62
X follows	0.43
X likes	0.38
X comments	0.43
X have photo with tag me	0.54
Followers	0.51
Common geotag	0.42
Common hashtag	0.36
X is a selebrity	0.57

friends” connection is stronger than the connection “i like j” ($E_{ij} \succ E'_{ij}$), therefore, in a further analysis of the relationship between i and j will be considered “school friends” and estimated in $p_{ij} = 0.4443$.

3.2. Weighted average

With knowledge of the reliability of sources, a combination of estimates can be obtained using a weighted average [25, 26]. Thus, the merging probabilistic relationship assessment will have the form $\widehat{p}_{ij} = p_{ij} \cdot w_{ij} + p'_{ij} \cdot w'_{ij}$, where p_{ij} and p'_{ij} is probabilistic estimates obtained using relationship information aboutE_{ij} and E'_{ij} respectively, w_{ij}, w'_{ij} is weights such that $w_{ij} + w'_{ij} = 1$.

At the same time, knowledge about the reliability of sources can be obtained in different ways, for example, by quantifying expert opinions or by obtaining estimates of the communication activity of users in social networks [27]. This approach will be explored further as part of further research. In addition, in the future it is planned to consider the possibility of applying the conjunctive combination rule [28], or use in a weighted average estimate of weights based on inverse dispersion.

4. Result

The approach based on highlighting the strongest connection is useful and convenient to use in case of differences in the probabilistic estimates of relationships. It is expected that this approach will give an optimal result when applied in the analysis of the social graph of the organization’s employees in the context of protection from social engineering attacks. However, the question of the applicability of this approach to other tasks in the analysis of social networks has not been studied. In addition, it may not be applicable in the case of approximately

equal estimates. Also, in [29] it is noted that the propagation of information in a social graph will be more difficult by the removal of a weak connection, which also confirms the need to verify proposed approach.

While the approach based on the weighted average value can be applied regardless of the values of probabilistic estimates. Nevertheless, it requires additional, more in-depth studies to find the optimal weights. It is also assumed that in the future these two approaches can be combined.

5. Conclusions

Thus, the article proposed approaches to the merging of probabilistic estimates of the relationship between users, based on the assumption that these probabilistic estimates are obtained by quantification. The theoretical significance of the results lies in the proposal of new approaches to the merging of probabilistic estimates of linguistic variables, the practical significance is seen in creating the basis for further analysis of the social graph of the organization's employees, in particular, for detecting the most critical trajectories of attack development or solving backtracking tasks of social engineering attacks.

Acknowledgments

The work was carried out as part of the project according to the state task SPIIRAS No. 0073-2019-0003, with financial support from the Russian Foundation for Basic Research, project No. 18-01-00626 — Methods for representation, truth estimates synthesis, and machine learning in algebraic Bayesian networks and related models of knowledge with uncertainty: probabilistic-logic approach and graph systems; Project No. 20-07-0083 — Digital twins and soft computing in social engineering attacks modelling and associated risks assessment.

References

[1] Sophos Whitepaper 2020, The impossible puzzle of cybersecurity: Results of an independent survey of 3,100 it managers commissioned by sophos, Sophos Whitepaper. URL: https://secure2.sophos.com/en-us/medialibrary/Gated-Assets/white-papers/sophos-impossible-puzzle-of-cybersecurity-wp.pdf, 2020. URL: https://secure2.sophos.com/en-us/medialibrary/Gated-Assets/white-papers/sophos-impossible-puzzle-of-cybersecurity-wp.pdf.

[2] Q. G. J. L. Q. Yan, F.R. Yu, Software-defined networking (sdn) and distrib-uted denial of service (ddos) attacks in cloud computing environments: A survey, some research issues, and challenges., IEEE communications surveys & tutorials 18 (2016, doi: 10.1109/COMST.2015.2487361) 602–622.

[3] G. B. V. B. J. G. C. F. G. R.-G. V. Elamaran, N. Arunkumar, Exploring dns, http, and icmp response time computations on brain signal/image databases using a packet sniffer tool., IEEE Access, (992018) 6 (2018, doi: 10.1109/ACCESS.2018.2870557) 59672–59678.

[4] M. R. V. Chang, Y. Kuo, Cloud computing adoption framework: A security framework for business clouds., Future Generation Computer Systems 57 (2016, doi: 10.1016/j.future.2015.09.031) 24–41.

[5] A. T. M.V. Abramov, T.V. Tulupyeva, Social Engineering Attacks: social networks and user security estimates, SUAI, St. Petersburg, 2018.

[6] Lenta2020, Telephone fraud affects one third of russians, Lenta.ru. URL: https://news.mail.ru/society/38435225/?frommail=1, 2020. URL: https://news.mail.ru/society/38435225/?frommail=1.

[7] Sputnik2020, Almost half of the inhabitants of lithuania were deceived by telephone and internet scammers, Sputnik, URL: https://www.kurier.lt/polovina-zhitelej-stalkivalas-s-elektronnymi-ili-telefonnymi-moshennikami/, 2020. URL: https://www.kurier.lt/polovina-zhitelej-stalkivalas-s-elektronnymi-ili-telefonnymi-moshennikami/.

[8] SecurityLab 2020, St. petersburg police caught three phishers, SecurityLab.Ru. News. URL: https://www.securitylab.ru/news/509364.php, 2020. URL: https://www.securitylab.ru/news/509364.php.

[9] RIA News 2020, The head of group-ib called the most common types of cybercrime in russia, Ria News. URL: https://ria.ru/20200617/1573066952.html, 2020. URL: https://ria.ru/20200617/1573066952.html.

[10] RIA 2020, Sberbank warned of new coronavirus schemes of fraudsters, Ria News. URL: https://yandex.ru/turbo/s/forbes.ru/newsroom/finansy-i-investicii/397727-sberbank-predupredil-o-novyh-shemah-moshennichestva-na-fone, 2020. URL: https://yandex.ru/turbo/s/forbes.ru/newsroom/finansy-i-investicii/397727-sberbank-predupredil-o-novyh-shemah-moshennichestva-na-fone.

[11] Kommersant 2020, Sberbank predicts economic losses from cybercrime in 2021 to 7 trillion rubles, Kommersant. URL: https://www.kommersant.ru/doc/4381088, 2020. URL: https://www.kommersant.ru/doc/4381088.

[12] A. S. A. T. M. A. R. U. A.A. Azarov, T.V. Tulupyeva, Social Engineering Attacks: The problems of analysis, Nauka, St. Petersburg, 2016.

[13] D. Katalkov, How social engineering opens the door to your organization for a hacker., Positive Research 2018. Practical Safety Research Digest., URL: https://www.ptsecurity.com/upload/corporate/ru-ru/analytics/Positive-Research-2018-rus.pdf, 2018. URL: https://www.ptsecurity.com/upload/corporate/ru-ru/analytics/Positive-Research-2018-rus.pdf.

[14] M. A. A. T. A.A. Korepanova, V.D. Oliseenko, Application of machine learning methods to the user accounts identification in two social networks, Computer tools in education 3 (2019) 38–43.

[15] T. T. A.A. Korepanova, User identification across different social networks through social circles, in: Information Security of Russian regions (ISRR-2019). XI St. Petersburg interregional conference., volume 3 of *Proceedings of the conference.*, SPOISY, St. Petersburg., 2019, pp. 442–443.

[16] R. H. T. M. Y. Yang, H. Yu, A fusion information embedding method for user identity matching across social networks., 2018 IEEE SmartWorld, Ubiqui-tous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innova-

tion (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) 18 (2018, doi: 10.1109/SmartWorld.2018.00340) 2030–2035.

[17] A. T. A.O. Khlobystova, M.V. Abramov, Identifying the most critical trajectory of the spread of a social engineering attack between two users., in: The Second International Scientific and Practical Conference "Fuzzy Technologies in the Industry – FTI 2018"., CEUR Workshop Proceedings, 2018, pp. 38–43.

[18] B. K. A.Y. Denzhakov, Methods and means of formalizing data in social networks., Actual problems of the humanities and natural sciences 12 (2010) 44–48.

[19] K. L. A.I.E. Hosni, Minimizing the influence of rumors during breaking news events in online social networks., Knowledge-Based Systems 18 (2019, doi: 10.1016/j.knosys.2019.105452).

[20] B. R. D.K. Srivastava, Words are important: A textual content-based identity resolution scheme across multiple online social networks., Knowledge-Based Systems 195 (2020, doi: 10.1016/j.knosys.2020.105624) 105624.

[21] A. T. A.O. Khlobystova, M.V. Abramov, Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users., in: D. V. E. B. D. I. M. Kotenko I., Badica C. (Ed.), Intelligent Distributed Computing XIII. IDC 2019., Studies in Computational Intelligence, Springer, Cham, 2019, doi: 10.1007/978-3-030-32258-8_32, pp. 272–277. doi:10.1007/978-3-030-32258-8_32.

[22] T. T. A. K. A.O. Khlobystova, A.G. Maksimov, An approach to quantification of relationship types between users based on the frequency of combinations of non-numeric evaluations., in: S. V. S. A. Kovalev S., Tarassov V. (Ed.), Proceedings of the Fourth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'19)., volume 1156 of *IITI 2019. Advances in Intelligent Systems and Computing*., Springer, Cham. doi: 10.1007/978-3-030-50097-9_21, 2020, pp. 206–213.

[23] L. Zadeh, The concept of a linguistic variable and its application to approximate reasoning., Learning systems and intelligent robots. Springer, Boston, MA (1974. doi: 10.1007/978-1-4684-2106-4_1) 1–10.

[24] L. Zadeh, Linguistic variables, approximate reasoning and dispositions., Medical Informatics 8 (1983) 173–186.

[25] K. W. J. Smith, A simple explanation of the forecast combination puzzle., Oxford Bulletin of Economics and Statistics 71 (2009, doi: 10.1111/j.1468-0084.2008.00541.x) 331–355.

[26] J. L. D. Li, W. Zeng, Fuzzy group decision-making based on variable weighted averaging operators., in: 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, Beijing. doi: 10.1109/FUZZ-IEEE.2014.6891632, 2014, pp. 1416–1421.

[27] S. Z. R.R. Tolstyakov, N.V. Zlobina, Research on social media use: theoretical and practical approaches., Bulletin Michurinsk State Agrarian University 4 (2016) 85–95.

[28] T. Denoeux, Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence., Artificial Intelligence 172 (2008, doi: 10.1016/j.artint.2007.05.008) 234–264.

[29] O. Kuznetsov, Models of activity propagation processes in network structures, in: XII All-Russian Meeting on Management, volume 3, Moscow, 2014, pp. 16–19.