Применение метода вероятностей альтернатив при построении оценок интенсивности взаимосвязей

А. О. Хлобыстова[1], М. В. Абрамов[2]
Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН),
Санкт-Петербургский государственный университет
Санкт-Петербург, Россия
[1]aok@dscs.pro, [2]mva@dscs.pro

Т. В. Тулупьева
Северо-Западный институт управления РАНХиГС,
Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН),
Санкт-Петербургский государственный университет
Санкт-Петербург, Россия
tvt@dscs.pro

***Аннотация*. Одной из наиболее актуальных проблем в сфере кибербезопасности является стабильный рост числа успешных социоинженерных атак. Для анализа защищённости от таких атак зачастую требуется уметь оценивать интенсивность взаимодействия пользователей в социальных сетях. Однако данная информация представляет собой некоторое множество лингвистических переменных и нуждается в квантификации. Целью настоящей статьи является предложение подхода к означиванию параметров модели для оценки вероятности успеха многоходовой социоинженерной атаки злоумышленника на пользователя. Для достижения данной цели впервые были предложены три различные модификации метода квантификации Н.В. Хованова, основывающиеся на предположении, что эксперты могли не только ранжировать лингвистические переменные, но и дать им некие примерные оценки, тем самым существенно расширив возможности для дальнейшей квантификации. Теоретическая значимость исследования заключается в предложении нового подхода получения вероятностных оценок по нечисловой информации. Практическая значимость результатов заключается в построении основы для дальнейшего использования при расчете оценок вероятности распространения многоходовых социоинженерных атак и анализе социального графа сотрудников организации. Все это закладывает фундамент для последующей диагностики информационных систем на предмет выявления уязвимостей к социоинженерным атакам, а также при решении задач социокомпьютинга.**

Ключевые слова: социинженерные атаки; квантификация; лингвистическе переменые; анализ социальных связей; социальный граф пользователей; вероятностные оценки

I. Введение

Одной из наиболее актуальных проблем в сфере кибербезопасности остаётся стабильный рост числа успешных несанкционированных атак [18] с применением прикладных психологических и аналитических приёмов скрытой мотивации пользователей информационных систем к нарушениям устоявшихся правил и политик в области информационной безопасности – социоинженерных атак [10],[11]. Так, по данным Лаборатории Касперского [1], около 90% утечек корпоративных данных происходит из-за ошибок человека, спровоцированных с помощью социальной инженерии, а не из-за технических проблем. Актуальность данной проблемы подтверждают и эксперты в области безопасности информационных систем [3], [6], [9]. Тем не менее, несмотря на разработанность тематики повышения технических аспектов безопасности систем [16], [17], область социоинженерных атак изучена в меньшей степени. Согласно [12], в большинстве случаев атакующие социоинженеры оперируют информацией, собранной из открытых источников, в том числе из социальных сетей. Вот почему важно учитывать такие данные при анализе защищённости пользователей информационных систем от социоинженерных атак.

Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2019-0003, при финансовой поддержке РФФИ, проекты №18-01-00626 и №20-07-00839

II. Релевантные работы

Методы извлечения и анализа данных из социальных сетей рассматривались в [2], [4], [7], [10], [14]. Однако, большая часть информации, получаемой из социальных сетей, характеризуется нечёткостью и неопределенностью в её измерении и вычислении, что делает её применение в математических моделях и дальнейшем автоматизированном анализе крайне затруднительным. Подходы к квантификации различных типов взаимоотношений, обозначаемых пользователями в социальных сетях, были предложены в [5], [20]. Числовые данные, полученные таким способом, дают возможность строить оценки вероятности прохождения злоумышленника по той или иной траектории, производить их сравнение и выявлять наиболее критичные. Тем не менее, полученные в [5], [20] оценки могут быть улучшены за счет агрегации сведений о большем числе параметров, влияющих на них.

III. Постановка цели и задач

При построении оценок защищенности пользователей информационных систем от многоходовых социоинженерных атак (при совершении которых задействуется цепочка пользователей) требуется знать оценки вероятностей распространения атаки злоумышленника от пользователя к пользователю. Такие оценки могут строиться на основе информации об интенсивности взаимодействия пользователей в социальных сетях, которые в свою очередь могут быть получены путём проведение социологического опроса.

Обозначим рассматриваемые в опросе связи (лингвистические переменные) через R_i. В таком случае результатами опроса будет не численная мера силы связи, а только порядки (“ < ”, “ > ”, “ ≈ ”) силы связи в зависимости от ее видов. К примеру, эксперт мог сказать, что $R_i > R_j$ – интенсивность связи «родственники» сильнее (“>”), чем связи «друзья по школе». Тогда мы можем ранжировать между собой все связи, на основании представления о них опрошенного респондента. Квантификация связей, полученных в результате опроса и основывающаяся на ранжировании связей респондентами, была рассмотрена в [5], [20] и опиралась на методы, предложенные исследовательским коллективом Н.В. Хованова [13], [15], [21]. Однако за счет агрегации сведений о большем числе параметров оценки могут быть улучшены. Таким образом, целью настоящей статьи является предложение подхода к означиванию параметров модели для оценки вероятности успеха многоходовой социоинженерной атаки злоумышленника на пользователя за счёт предложения модификаций метода квантификации лингвистических переменных.

Теоретическая значимость исследования заключается в предложении нового подхода получения вероятностных оценок по нечисловой информации. Практическая значимость результатов заключается в построении основы для дальнейшего использования при расчете оценок вероятности распространения многоходовых социоинженерных атак и анализе социального графа сотрудников организации. Все это закладывает фундамент для последующей диагностики информационных систем на предмет выявления уязвимостей к социоинженерным атакам, а также при решении задач социокомпьютинга.

IV. Теоретическа основа

Согласно методам, предложенным в [13], [15], [21], при невозможности получения точных числовых оценок, но наличии совокупности сравнительных суждений о расположении оцениваемых объектов по отношению друг к другу может быть использован метод построения оценок вероятностей альтернатив по нечисловой, неточной и неполной информации. Данный метод в свою очередь основывается на байесовской модели рандомизации неопределённости выбора вектора оценок вероятности из множества всех допустимых векторов. Опишем применение данного метода к рассматриваемым в данной статье лингвистическим переменным – взаимосвязям между пользователями.

Предположим, что требуется получить оценки интенсивности по m различным видам связей. Пусть уже проведён опрос экспертов и на основе их ответов получено множество пар $\{(O_i, F_i)\}_{1\le i\le n}$, где $O_i = \left\langle R_{j_1}, R_{j_2}, ..., R_{j_k} \right\rangle_{1\le k\le m}$ – упорядоченный набор ранжированных связей (порядок связей), F_i — число экспертов, расположивших связи в данном порядке (порядковая частотность). Заметим, что мощность $R_{j_k, 1\le k\le m}: 1 \le \left|R_{j_k}\right| \le m$, то есть, если $\left|R_{j_k}\right| > 1$, то R_{j_k} обозначает не одну связь, а несколько, которые были оценены экспертами, как примерно равные. Например, если эксперт использовал порядок “ ≈ ” в отношении связей R_1, R_2, R_3 (т.е. оценил $R_1 \approx R_2 \approx R_3$), то в упорядоченном наборе O_i они будут рассматриваться как одна лингвистическая переменная (R_{j_k}), имеющая мощность 3 ($\left|R_{j_k}\right| = 3$).

Для применения метода Н.В. Хованова требуется ввести шкалу возможных вероятностных значений: $\{p_0 = 0, p_1 = 1/n, ..., p_{N-1} = (N-1)/N, p_N = 1\}$. После чего посчитать математическое ожидание каждой связи в каждом порядке. Например, пусть нужно оценить 3 связи: $O = \langle R_1, R_2, R_3 \rangle$, $|R_i| = 1, i = 1,2,3$, $N = 3$, множество возможных вероятностных значений: $\left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\}$. В таком случае возможно 4 варианта различного распределения R_i (таблица).

ТАБЛИЦА I ВАРИАНТЫ РАСПРЕДЕЛЕНИЯ R_i

	0	1/3	2/3	1
1	R_1	R_2	R_3	
2		R_1	R_2	R_3
3	R_1		R_2	R_3
4	R_1	R_2		R_3

$\mathrm{E}[R_1] = 0\cdot 3/4 + 1/3\cdot 1/4 = 1/12$, $\mathrm{E}[R_2] = 1/3\cdot 2/4 + 2/3\cdot 2/4 = 1/2$, $\mathrm{E}[R_3] = 2/3\cdot 1/4 + 1\cdot 3/4 = 11/12$. Посчитав математическое ожидание для каждого порядка из множества $\{(O_i, F_i)\}_{1\le i\le n}$, найдём:

$$\forall i:\ \mathrm{EE}\left[R_i\right] = \sum_{k=1}^{n} \frac{F_k}{r} \cdot \mathrm{E}\left[R_i^k\right], 1 \le i \le m, \qquad (1)$$

где r – число всех экспертов, n – число различных порядков. Таким образом, получили $\mathrm{EE}\left[R_i\right], 1 \le i \le m$ – агрегированные оценки экспертов. Рассмотрим модификации данного метода.

V. Предлагаемые модификации

A. Задание границ

Пусть в ходе опроса экспертам предлагалось не только ранжировать связи, но и интуитивно оценить силу каждого вида связи по шкале от 0 до 1. Т.е. эксперт мог дать оценку p_i связи R_i и p_j связи R_j (например, оценить связь «родственники» в 0.87, а связь «друзья по школе» в 0.74). В таком случае на шкалу возможных вероятностных значений могут быть наложены дополнительные ограничения по каждому из видов связи, и как следствие будут получены новые вероятностные оценки интенсивности связей.

Предположим, что нам даны результаты опроса экспертов с их интуитивными оценками каждого из типов взаимоотношений. Первым шагом выполним их предобработку, а именно произведём отчистку данных от выбросов и аномальных значений. После чего сопоставим каждому виду связи минимальное покрывающее множество возможных вероятностных значений. Если связи R_i соответствует множество оценок $\{p_1,...,p_r\}$, где r – число всех экспертов, тогда минимальное покрывающее множество вероятностных значений будет иметь вид $P^i=\left\{p^i_0, p^i_1,\ldots, p^i{}_{S-1}, p^i{}_S\right\}$. То есть P_i – дискретное множество с фиксированным расстоянием между элементами $\forall k: 0\le k\le S \;\; p^i{}_{k+1}-p^i{}_k=1/N$ и при этом $p^i_0\le \min\{p_1,...,p_r\}$, $p^i_S\ge \max\{p_1,...,p_r\}$. Заметим, что в данном случае N – константа – выбранное деление шкалы вероятностных значений, которое совпадает для всех связей R_i, а S – переменный индекс (может не совпадать для различных связей), вычисляемый в процессе формирования множества P_i. При таком подходе для каждого порядка будут получены новые значения математического ожидания и, как следствие, новые вероятностные оценки силы связей по агрегированным оценкам экспертов.

B. Ранжирование по группам

Пусть верны предположения, сделанные в предыдущем разделе. Разделим m связей по группам. Для этого воспользуемся статистическим программным обеспечением и произведём корреляционный анализ оценок, полученных в результате опроса респондентов. В зависимости от полученных коэффициентов корреляции определим значение $b\,(0.6\le b\le 1)$. Если значение коэффициента корреляции, соответствующее связи между R_i и R_j, будет больше или равно заданному b, то R_i и R_j будут определены в одну группу.

После чего произведём ранжирование полученных групп. Одним из способов такого ранжирования может являться сравнение средних значений. Затем разобьём шкалу возможных вероятностных значений на части, пропорциональные числу связей в полученных группах. Поясним это на примере.

Пусть число групп равно 4, а число связей – 10, связи по группам распределены следующим образом: $G_1=\{R_3\}, G_2=\{R_1,R_2,R_7,R_8,R_{10}\}, G_3=\{R_5,R_6\}, G_4=\{R_4,R_9\}$, где R_i обозначает связь, а G_j – группу. Для дальнейшего применения метода Н.В. Хованова введём шкалу возможных вероятностных значений: $\left\{0,\frac{1}{N},...,\frac{N-1}{N},1\right\}$. Разобьём отрезок $[0,1]$ на 10 (общее число связей) равных частей (т.е. $N=10$). Группе G_1 будет соответствовать $1/10$ такого отрезка, взятая с его начала: $[0,0.1]$. Рассмотрим крайнюю левую точку полученного промежутка и сопоставим ей меньшее или равное значение из заданной шкалы возможных вероятностных значений – $1/T$. Аналогично крайней правой точке сопоставим большее или равное значение, принадлежащее вышеупомянутой шкале – $1/S$ (значения T и S – отличны для разных групп). Таким образом, $G_1=\{R_3\}$ будет соответствовать шкала: $\left\{0,...,\frac{1}{S}\right\}, \frac{1}{S}\le 0.1$. Аналогично для $G_2=\{R_1,R_2,R_7,R_8,R_{10}\}$: $\left\{\frac{1}{T},...,\frac{1}{S}\right\}$, где $\frac{1}{T}\le 0.1, \frac{1}{S}\ge 0.6$ и т.д.. Вследствие чего каждой группе будет соответствовать своё множество возможных вероятностных значений. Затем применим метод Н.В. Хованова, но уже внутри полученных групп и на сопоставленных им шкалах.

C. Вторичные опрос

Ещё одним способом уточнения получения числовых оценок вероятности может являться проведение дополнительного опроса. Такой опрос может быть проведён как по отношению к уже полученным результатам, так и независимо от них: по прошествии определённого срока и опросе тех же самых экспертов. В обоих случаях будем предполагать, что эксперты уже интуитивно ранжировали или оценили силу каждого вида связи по шкале от 0 до 1.

«Постопрос»

После чего могла быть произведена квантификация полученных результатов одним из предложенных выше методов. Затем переменные были упорядочены, согласно полученным числовым оценкам. Экспертам (в данном случае это могут быть как те, кто участвовал в опросе в первый раз, так и новые респонденты) предлагается изучить упорядоченные переменные и в случае несогласия с расположением некоторых из них указать их предполагаемое место в данном ряду.

В случае повторяющегося несогласия у определённого количества экспертов (например, у более, чем 50%) выполнить повторную процедуру квантификации, но уже с фиксированным положением некоторых переменных при выборе вектора оценок вероятности из множества всех допустимых векторов.

Также, данный подход является одним из вариантов верификации полученных результатов. И может быть использован для сравнения предлагаемых подходов. Однако

в данном случае при проведении «постопроса» должны быть сформированы независимые группы экспертов. Число групп совпадает с числом сравниваемых подходов. А сами эксперты в группах обладают примерно равным характеристиками, т.е. распределение по полу, возрасту и другим характеристикам совпадает во всех группах.

«Предопрос»

Ещё один вариант проведения вторичного опроса – предложить пройти тот же самый опрос по прошествии определённого количества времени экспертам, которые проходили его и в первый раз. В таком случае появляется возможность применения методов статистической обработки информации к двум выборкам. В ходе анализа которых могут быть выявлены переменные, сохранившие своё положение, и которые при дальнейшей квантификации на шаге выбора вектора оценок могут быть зафиксированы.

VI. Заключение

В ходе исследования был описан подход к означиванию параметров модели для оценки вероятности успеха многоходовой социоинженерной атаки злоумышленника на пользователя, а также предложены три различные его модификации. Для улучшения получаемых числовых оценок было сделано предположение, что эксперты могли не только ранжировать лингвистические переменные, но и дать им некие примерные оценки. Тем самым позволив наложить на шкалу возможных вероятностных значений дополнительные условия, например, такие как задание вероятностных границ или разделение связей на группы. Для применения третьего подхода предполагается проведение повторного опроса, ожидается, что данный подход даст наиболее точные числовые оценки, однако к его недостатку следует отнести трудность реализации на практике.

Результаты исследования могут быть использованы при построении и анализе социального графа сотрудников организации. Такой граф может быть ассоциирован с данными об активах информационной системы и сравнительной оценкой уровня их критичности [8], [19] для последующей диагностики организации на воздействие социоинженерных атак и принятия решений о мерах, снижающих риски реализации атак. Кроме того, результаты могут быть применимы и при решении задач социокомпьютинга. В дальнейшем планируется проведение эмпирического исследования на предмет сравнения оценок, получаемых при использовании исходного метода и трёх его модификаций, а также последующего выявления наиболее эффективного, правдоподобного и доступного к применению среди них.

Список литературы

[1] «Лаборатория Касперского»: 9 из 10 утечек данных из облаков происходит из-за человеческого фактора [Электронный ресурс] Лаборатории Касперского. О нас. Новости. 2019. //URL: https://www.kaspersky.ru/about/press-releases/2019_laboratoriya-kasperskogo-9-iz-10-utechek-dannyh (дата обращения: 18.02.2020).

[2] Bilal M., Gani A., Lali M.I.U., Marjani M., Malik N.Social Profiling: A Review, Taxonomy, and Challenges //Cyberpsychology, Behavior, and Social Networking. 2019. Vol. 22. №. 7. Pp. 433–450.

[3] Corradini I., Nardelli E. Social Engineering and the Value of Data: The Need of Specific Awareness Programs //International Conference on Applied Human Factors and Ergonomics. Springer, Cham, 2019. Pp. 59–65.

[4] Jang B., Yoon J. Characteristics analysis of data from news and social network services // IEEE Access. 2018. Vol. 6. Pp. 18061–18073.

[5] Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft Estimates for Social Engineering Attack Propagation Probabilities Depending on Interaction Rates Among Instagram Users // International Symposium on Intelligent and Distributed Computing. Springer, Cham, 2019. Pp. 272–277.

[6] Patel P., Kannoorpatti K., Shanmugam B., Azam S., Yeo K. A theoretical review of social media usage by cyber-criminals //2017 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2017. Pp. 1–6.

[7] Sambir A., Yakovyna V., Seniv M. Recruiting software architecture using user generated data // 2017 XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), Polyana, Ukraine, 20–23 Apr 2017 / IEEE, 2017. Pp. 161–163.

[8] Shakibazad M., Rashidi A.J. New method for assets sensitivity calculation and technical risks assessment in the information systems //IET Information Security. 2019. Vol. 14. №. 1. Pp. 133–145.

[9] The Human Factor 2018 Report [Electronic resource] Proofpoint. Threat center //URL: https://www.proofpoint.com/us/human-factor-2018 (last available: 15.03.2020)

[10] Абрамов М.В., Тулупьева Т.В., Тулупьев А.Л. Социоинженерные атаки: социальные сети и оценки защищенности пользователей. СПб.: ГУАП, 2018. 266 с.

[11] Азаров А.А., Тулупьева Т.В., Суворова А.В., Тулупьев А.Л., Абрамов М.В., Юсупов Р.М. Социоинженерные атаки. Проблемы анализа. СПб.: Наука, 2016. 349 с.

[12] Каледина А. В ЦБ обеспокоены ростом киберпреступлений на основе методов социальной инженерии [Электронный ресурс] Известия //URL: https://iz.ru/897320/anna-kaledina/nu-i-gadzhety-rossiiane-stali-samoi-legkoi-dobychei-dlia-kibermoshennikov (дата обращения: 16.02.2020).

[13] Колесников Г.И., Хованов Н.В., Юдаева М.С. Применение метода квантификации нечисловых оценок вероятности для выбора оптимального портфеля ценных бумаг // Вестник Санкт-Петербургского университета. Экономика. 2007. №. 3. С. 58–68.

[14] Корепанова А.А., Абрамов М.В., Тулупьева Т.В. Идентификация аккаунтов пользователей в социальных сетях «Вконтакте» и «Одноклассники» //Семнадцатая Национальная конференция по искусственному интеллекту с международным участием. КИИ–2019. Ульяновск, 21–25 окт. 2019. Т. 2. Ульяновск: УлГТУ, 2019. С. 153–163.

[15] Корников В.В., Хованов Н.В., Юдаева М.С. Многокритериальная классификация в условиях дефицита числовой информации //Труды Карельского научного центра Российской академии наук. 2012. №. 5. С. 38–43.

[16] Котенко И.В., Кулешов А.А., Ушаков И.А. Система сбора, хранения и обработки информации и событий безопасности на основе средств Elastic Stack // Труды СПИИРАН. 2017. Т. 54. №. 0. С. 5–34.

[17] Осипов В.Ю., Воробьев В.И., Левоневский Д.К. Проблемы защиты от ложной информации в компьютерных сетях // Труды СПИИРАН. 2017. Т. 4. №. 53. С. 97–117.

[18] Утечки данных. Россия. 2018 год [Электронный ресурс] InfoWatch. Ресурсы. Аналитические отчёты. 2019 //URL: https://www.infowatch.ru/analytics/reports/russia2018 (дата обращения: 18.01.2020).

[19] Федорченко А.В., Дойникова Е.В., Котенко И.В. Автоматизированное определение активов и оценка их критичности для анализа защищенности информационных систем // Труды СПИИРАН. 2019. Т. 18. №. 5. С. 1182–1211.

[20] Хлобыстова А.О., Абрамов М.В., Тулупьева Т.В., Тулупьев А.Л. Социальное влияние на пользователя в социальной сети: типы связей в оценке поведенческих рисков, связанных с социоинженерными атаками // Управленческое консультирование. 2019. № 3. С. 104–117.

[21] Хованов Н.В., Федотов Ю.В. Модели учета неопределенности при построении сводных показателей эффективности деятельности сложных производственных систем // Научные доклады. 2006. №. 28R-2006. С. 37–87.