UDC 004.912

The analysis of cybersecurity subject area terms based on the information diffusion model

Dmytro Lande[1,2], Olexiy Novikov[1], Dmytro Manko[2]

[1] *National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”*
[2] *Institute for Information Recording of NAS of Ukraine*

Abstract

This research describes a comparison of the information diffusion model, built on the basis of cellular automata with the real statistics the dynamics of the use of terms from the field of cybersecurity in the information flows of the Internet. The information diffusion model is used with different parameters of the intensity of the information propagation. The cross-correlation of dynamics of the dissemination of new information in the model with the dynamics of the occurrence of concepts of the real subject area has been calculated. A high correlation dynamic of the terms occurrence with the dynamics given by the model at the selected parameters is shown. The research results allow fixing the model parameters that can further perform forecasting. The advantage of the information diffusion model based on cellular automata is the simplicity and clarity of a small number of parameters, and the ability to change them in accordance with the data on the actual occurrence of special terminology in information flows. The dynamics of the information diffusion model under various parameters allows us to determine centroids for the subsequent clustering of domain terms.

Keywords: Modeling, Subject domain, Cross-correlation, Cellular automata, Information diffusion, Cyber security datasets

Introduction

Among the many methods of modeling the dissemination of information in computer (including social) networks [1-3]. While most such models are currently modeled by analytical methods using differential equations, the most obvious implementations of these approaches are based on the theory of cellular automata [4], which are supplemented with analytical calculations. Some information dissemination models are based on infection models, such as SIS, SIR or SIRS. SIS "susceptible — infected — susceptible" model (S for the number of susceptible, I for the number of infectious) is applicable to the analysis of the spread of diseases to which immunity is not developed. It is described by the following system of equations:

$$\frac{dS(t)}{dt} = -\beta\frac{S(t)I(t)}{N} + \gamma I(t),$$
$$\frac{dI(t)}{dt} = \beta\frac{S(t)I(t)}{N} - \gamma I(t). \qquad (1)$$

where

- $S(t)$ — the number of susceptible individuals at time *t*;
- $I(t)$ – the number of infected individuals at time *t*;
- N – population size;
- β — coefficient of the intensity of contacts of individuals with subsequent infection;
- γ — coefficient of the intensity of the recovery rate of infected individuals

In the second case, the SIR model consists of three compartments: S for the number of **S**usceptible, I for the number of **I**nfectious, and R for the number of **R**ecovered, (or deceased, or immune) individuals.

The SIR model has gained popularity due to its ease of construction, implementation, and use. Its application allows you to accurately model influenza and other diseases in large cities, enter new parameters, and analyze different scenarios.

In the third, even more realistic case, the probability of loss of immunity in previously infected individuals is assumed. The SIRS principle - "susceptible – infected – recovered – susceptible", is a model for describing the dynamics of diseases with temporary immunity (recovered individuals become susceptible again over time).

Goals

In our work, we consider a model of information dissemination, in which a separate element of the system of cellular automata can be in a state of readiness for the perception of new information (susceptible), possession with the possibility of dissemination of information (infected), possession of information without the possibility of dissemination (analogue - recovered) and again in the state of forgetting information and readiness to perceive new information (again - susceptible). The purpose of this paper is to compare the diffusion information model, based on the concept of cellular automata discussed below, and the actual the process of appearance of domain terms (cybersecurity) in information flows. For this purpose, the SIRS model is built, which is used for various parameters of the intensity of the information dissemination. Then the cross-correlation of the information dissemination dynamics that the model provides is calculated with the real dynamics of the dissemination of the particular terms. For each the determined domain term, the model parameters for which the cross-correlation values are sufficiently large are fixed, and the model parameters that allow for further forecasting are likewise fixed.

Cellular automata

A cellular automaton is a discrete dynamic system, a collection of identical cells that are interconnected in a certain way. All cells form a network of cellular automata. The state of each cell is determined by the state of cells in its local neighborhood or its “nearest neighbors”. The state of the *j*-th cellular automaton at a time $t+1$ is thus determined as follows: $y_j(t+1)=F\left(y_j(t),O(j),t\right)$, where F is a rule that can be expressed, for example, in the language of Boolean algebra. In many problems, it is considered that the element relates to its closest neighbors, i.e. $y_j(t)\in O(j)$, in this case, the formula is simplified: $y_j(t+1)=F\left(O(j),t\right)$.

Cellular automata in the traditional sense satisfy such rules:

- the values of all cells are changed simultaneously (the unit of measurement is a clock cycle).
- the network of cellular automata is homogeneous, i.e. the rules for changing states for all cells are the same;
- the cell can only be affected by cells from its local neighborhood – the set of cell states is finite.

In the case of a two-dimensional lattice, whose elements are squares, the nearest neighbors entering the neighborhood of the element y_{ij} can be considered either only the elements located up-down and left-right from it (the so-called von Neumann neighborhood: $y_{i-1,j}, y_{i,j-1}, y_{i,j+1}, y_{i+1,j}$), or diagonal elements added to them (the so-called “Moore's neighborhood”: $y_{i-1,j-1}, y_{i-1,j}, y_{i-1,j+1}$, $y_{i,j-1}, y_{i,j}, y_{i,j+1}$, $y_{i+1,j-1}, y_{i+1,j}, y_{i+1,j+1}$). This allows us to determine the overall ratio of the cell value at step t+1 compared to step t [4]:

$$y_{i,j}(t)=F(y_{i-1,j-1}(t),y_{i-1,j}(t),y_{i-1,j+1}(t),y_{i,j-1}(t),y_{i,j}(t),y_{i,j+1}(t),$$
$$,y_{i+1,j-1}(t),y_{i+1,j}(t),y_{i+1,j+1}(t)).$$

The SIRS Model

The information diffusion model is two-dimensional, thus the whole system of cellular automata for this case is described by a two-dimensional array. In the case of a two-dimensional lattice whose elements are squares, the nearest neighbors that are around the element can be considered or only elements located up and down and left and right of it and added to it and diagonal elements (Moore's vicinity).

In the framework of the model of information diffusion, which refers to the dissemination of information stories in the information space, probabilisty rules of news dissemination are applied. Such a model is constructed, and is given the following semantic meaning: a system of square cellular automata (Moore's model) with four cellular states is considered:

1 – the latest news (black cell);

2 - the news is known, but it is in active state (gray cell);

3 - the cell has no information (the cell is white, the information has not reached or has been forgotten).

Rules

The model of information diffusion provides the following rules for the development of the information plot related to the news (Fig. 1):

1) initially, the whole field consists of white cells except for one - black, which was the first to "receive" the news;

2) the white cell can be repainted only in black or remain white (it can receive news or remain "in ignorance");

3) the white cell is repainted if the condition is met: $C \cdot pm > 1$, where p is a pseudo-random variable ($0 < p < 1$), m is the number of black cells in the vicinity, C is a constant (C = 1,5 at m = 1; C = 1 at $m \neq 1$);

4) if the cell is black, and around it black and gray ($s > x,\ x$ is the number of black and gray cells, x is a given constant,, the parameter of "random access memory") then it is repainted in gray (news becomes obsolete, but is stored as information);

5) if the cell is gray, and around it only black and gray (($s > y,\ y$ –a constant, a parameter of "archive" memory), it is repainted in white (forgetting information when they are well known).

The MATLAB software has been used to implement the model [5].

Deviations from the nature of a smooth "burst" become inherent in information flows that, in particular, correspond to the topic of cyber security":

- rapid termination of "undesirable" information plot (*S*-effect);
- stretching the period of raising the information plot (*L*-effect) with the "desired" administration topic.

These deviations are manifested in the model in the case when the parameters of the rules change, which, in particular, determine the behavior of the model of information diffusion, corresponding to some life observations. If you compare the black cells of the model (real-time message) finding the message in RAM, and gray - finding the message in the archive memory, the *S*- or *L*-effects will correspond to the ratio of the time the message is in RAM or archive memory governed by the parameters *x* and *y* of rules 4 and 5.

At values of parameters *x*=*y*=8 the model corresponds to the natural dynamics of an information message development outside the information reservation - its graph takes the form of a close to symmetric curve.

When saving the RAM parameter (*x* = 8) and reducing the archive memory parameter *y*, (up to *y* = 2), gray cells are more often released from information, then more intensely perceive previously forgotten information, repainting in black, ie there is an effect of "pushing" a new message (corresponding to the scenario of information impact) -which is the *L*-effect.

On the other hand, while saving the RAM parameter (*x* = 8) and reducing the archive memory parameter (to *x*=2) is a quick "forgetting" of the message, which does not correspond to the information impact, and moving its main part to the archive – which is *S*-effect.

On the other hand, while saving the RAM parameter (*x*=8). Typical dependences of the number of cells (sequence of the number of cells of the same type), which are in different states (number of gray cells -x_g, white - x_w and black - x_b from the evolutionary step) as a result of analytical approximation are expressed by formulas:

$$x_g = \frac{1}{1+e^{-\alpha(t-\tau_1)}}; \qquad x_b = \frac{1}{1+e^{-\beta(t-\tau_2)}} - \frac{1}{1+e^{-\alpha(t-\tau_1)}}$$

$$x_w = 1 - \frac{1}{1+e^{-\beta(t-\tau_2)}};$$

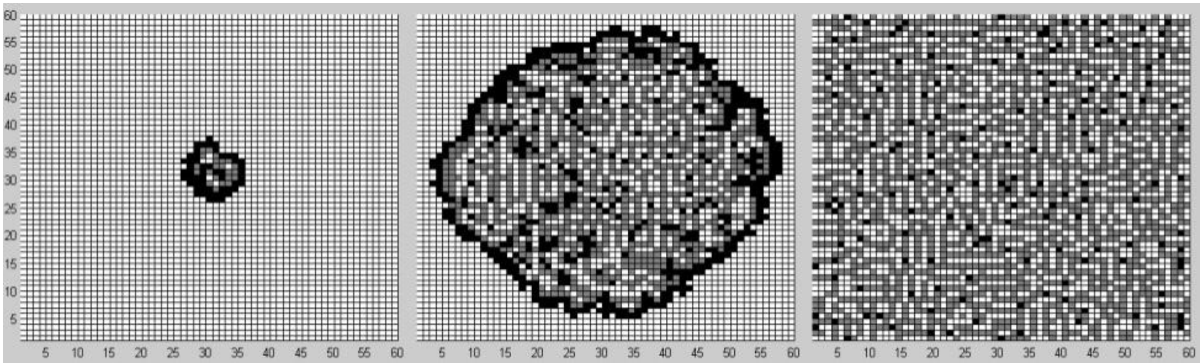

Fig. 1. The intermediate states of cellular automata systems at various stages

Basic profiles of the dynamics of information plots, corresponding to the values of x=y= 8 in the rules 4 and 5 of the model, were obtained at the values of the parameters $\boldsymbol{\alpha = 0.15}$, $\boldsymbol{\beta = 0,25}$.

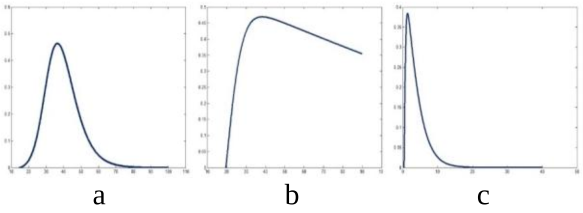

Fig. 2. Dynamics of the number of cells in the state of latest news:
a – the typical dynamics (α = 0,15, β =0,25);
b – stretching the period of relevance of information (α = 0,01, β = 0,25);
c – immediate cessation of dissemination of (α = 0,15, β = 1,5) information.

Sources of Information and Timelines

About 6,000 web sources (active part of the Russian and Ukrainian segments of the Web space for publications on cyber-attacks) were analyzed in March-April 2022 from the content monitoring system Infostream (http://infostream.ua). This request has been processed:

cyberattack | cyberattack | (cyber ~ attack) | (cyber attack)

A total of **9495** documents were received on request, the dynamics of publications on request is shown in Fig. 3.

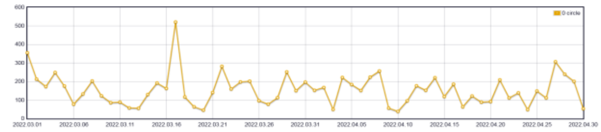

Fig. 3. Dynamics of the number of publications corresponding to the request, by days

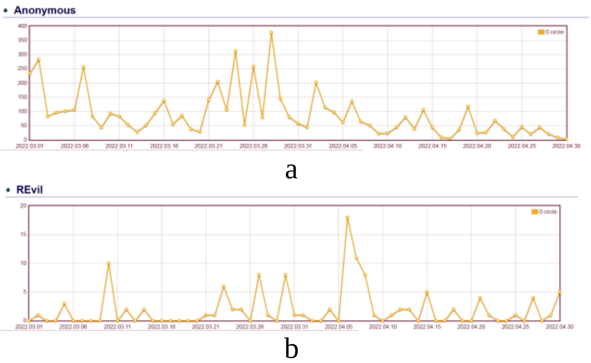

Fig. 4. Dynamics of the number of publications corresponding to the subjects (*a* - Anonymous, *b* - REvil)

Based on the linguistic and statistical analysis of the information flows studied in the creation of these reports, a large number of subjects of cybersecurity have been identified. In Fig. 4 shows the dynamics of the number of publications corresponding to two of them Anonymous and REvil.

The most popular terms that correspond to the subjects of cybersecurity in the online media are the following:

№	Subject	Description
1	Anonymous	“Anonymous” hacker group
2	REvil	Cyber extortion group “Revil”
3	Killnet	“Killnet” hacker group
4	The Infraud Organization	Hacker group “The Infraud Organization”
5	Mandiant	“Mandiant” Research Group
6	Lurk	“Lurk” hacker group
7	CrowdStrike	Cybersecurity company “CrowdStrike”
8	Check Point Research	Research company “Check Point Research”
9	Lazarus Group	Hacker group “Lazarus”
10	ESET	ESET research company

Correlation between the dynamics of the model and real data

To assess the adequacy of the proposed model, a cross-correlation of the dynamics of mortality in the model (the number of black cells per clock cycle) and real data on the dynamics of the emergence of terms corresponding to the subjects of cybersecurity, is calculated [6]. In this case, the simulation results depend on the parameters α and β. For each pair of parameter values from the range $(0.01 \le \alpha < 0.2;\ \ 0.1 \le \beta < 2.0)$, the dynamics of the appearance of black cells was calculated. For each of these dynamics vectors, a cross-correlation was calculated with the dynamics of mortality in the selected country. The maximum cross-correlation and corresponding parameter values were selected for each country α and β (Table 1).

The set of maximal cross-correlations between the obtained vectors is calculated, and the corresponding correlation matrix is formed with the elements in the notation of the formula (1):

$$a_{ij}(m) = \max_{m} \frac{\sum\limits_{k=1}^{n-m} w_{k+m}^{i} w_{k}^{j}}{\sqrt{\sum\limits_{k=m+1}^{n} \left(w_{k}^{i}\right)^{2}} \sqrt{\sum\limits_{k=1}^{n-m} \left(w_{k}^{j}\right)^{2}}}. \qquad (1)$$

The max function is used for the reasons that processes that are similar in nature may have similar dynamic behavior, but possibly with a time shift [8].

In table 1 there are some examples of the dynamics of the appearance of black cells in the model under study, the dynamics of real mortality processes, and the surface of cross-correlations with different values of α and β.

As can be seen, for almost every considered object, the maximum cross-correlation value exceeds 0.5, which indicates the adequacy of the model and its predictive capabilities.

Table 1. Values of the maximum cross-correlation of model parameters for some terms

Subject	Parameter α	Parameter β	The calculated maximum cross-correlation
Anonymous	0.13	1.4	0.582
REvil	0.13	1.4	0.516
Killnet	0.13	1.5	0.506
The Infraud Organization	0.14	1.5	0.746
Mandiant	0.15	0.3	0.643
Lurk	0.12	1.1	0.508
CrowdStrike	0.13	1.4	0.684
Check Point Research	0.16	0.4	0.628
Lazarus Group	0.13	1.2	0.548
ESET	0.17	0.3	0.515

Accordingly, subjects for which maximum values for cross-correlations are reached at close parameter values can be assigned to one cluster. The example shows that the selected subjects of cybersecurity were divided into two distinct groups - hacker teams and research companies (Fig. 5).

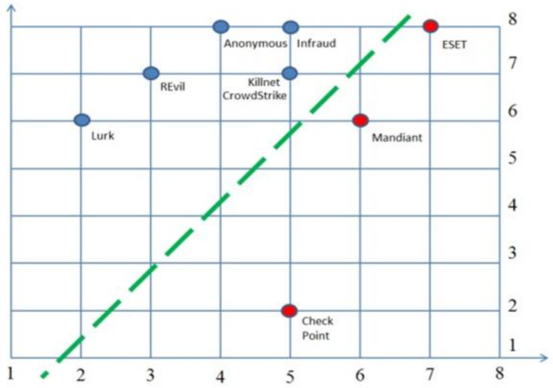

Fig. 4. Subjects of cybersecurity in the space of operative and archive memory parameters

Conclusions

The purpose of this paper is to compare the SIRS model, based on the concept of cellular automata discussed above, and the actual daily the dynamics of the emergence of various terms on the selected issue. For this purpose, the SIRS model is built, which is launched for various parameters. Then we calculate the cross-correlation of the dissemination of information rate dynamics that the model gives with the real mortality rate dynamics for different countries. For each term the model parameters for which the cross-correlation values are sufficiently large are fixed, and likewise, the model parameters are fixed that allow for further forecasting. In addition, the correlation of the dynamics of different terms with a model with the same parameters can be considered as a basis for conducting cluster analysis.

The study showed that it is possible to classify subjects according to the parameters of the proposed model of cellular automata (in the simplest case, α and β). This is the case when the model of cellular automata is not only descriptive and qualitative analysis, but can be used for real calculations. High values of correlation of model values and real dynamics of appearance of terms in an information stream testify to adequacy of model.

The advantage of the SIRS model is the simplicity and clarity of its parameters, and the ability to change them. It is also possible to convert the model to an analytical one, which will make it possible to change the number of "neighbors" during local interactions. Also, there are the following advantages: a small number of parameters, a relatively low dimension of the vectors of parameters, a reliable mathematical basis for correlation analysis, objectivity – a reliable aggregator of data answers for the "purity" of the data, the application of standard software tools, and its relative simplicity (ready-made software systems such as Matlab, Excel, R language, etc., can be applied).

The disadvantage of the model is that it only takes into account local interactions. At the same time, it can be expanded.

References

[1] Bhargava S.C., Kumar A., Mukherjee A. A stochastic cellular automata model of innovation diffusion // Technological Forecasting and Social Change. – 1993. – Vol. 44; Iss. 1.

[2] Minglei Fu, Hongbo Yang, Jun Feng, Wen Guo, Zichun Le, Dmytro Lande, Dmytro Manko. Preferential information dynamics model for online social networks (2019). Physica A: Statistical Mechanics and its Applications. Vol. 506. pp. 993-1005. DOI: 10.1016/j.physa.2018.05.017

[3] Ланде Д.В., Додонов В.А. Модель розповсюдження інформації з урахуванням поняття сприйняття і пам'яті // Авіаційна та екстремальна психологія у контексті технологічних досягнень: збірник наукових праць / – Київ: Аграр Медіа Груп, 2017. – С. 148-153. (http://dwl.kiev.ua/art/psy/index.html).

[4] Harko, Tiberiu; Lobo, Francisco S. N.; Mak, M.K. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. Applied Mathematics and Computation (2014). 236: 184–194. DOI: 10.1016/j.amc.2014.03.030

[5] Greenhalgh D. & Moneim I.A. SIRS Epidemic Model and Simulations Using Different Types of Seasonal Contact Rate. Systems Analysis Modeling Simulation (2003). Vol. 43. Iss. 5. PP. 573-600. DOI: 10.1080/ 023929021000008813.

[6] S. Wolfram “New Kind of Science”. Wolfram Media (2002). – 1259 p. ISBN: 1579550088, 9781579550080.