СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 004.94

ПОСТРОЕНИЕ МОДЕЛИ РАСПРОСТРАНЕНИЯ ВИРУСА В КОМПЬЮТЕРНОЙ СЕТИ НА ОСНОВЕ СРАВНЕНИЯ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ С ЭМПИРИЧЕСКИМИ ДАННЫМИ

Еремеева Н.И.
Димитровградский инженерно-технологический институт – филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ», г. Димитровград

Поступила в редакцию 21.11.2022, после переработки 20.12.2022.

В статье на основе SI-модели построена иерархия четырех математических моделей, описывающих распространение компьютерного вируса в сети при отсутствии возможности лечения зараженных хостов. Каждая следующая модель отличается от предыдущей введением в рассмотрение нового фактора. Для построенных моделей проведено сравнение результатов численного моделирования с экспериментальными данными (статистическими данными, опубликованными в 2002 году Д. Муром, К. Шенноном и Д. Брауном). На основе построенных моделей с использованием некоторых элементов PSIDR-модели, создана модель, описывающая распространение компьютерного вируса в сети при наличии возможности лечения зараженных хостов, а также проведен численный эксперимент, позволяющий проследить динамику численности всех групп, участвующих в эпидемиологическом процессе.

Ключевые слова: математическое моделирование, системы дифференциальных уравнений, распространения вируса в компьютерной сети.

Вестник ТвГУ. Серия: Прикладная математика. 2022. № 4. С. 39–52.
https://doi.org/10.26456/vtpmk647

Введение

Сетевые компьютерные технологии все более прочно входят в нашу жизнь и находят применение во многих сферах человеческой деятельности. Но наряду с очевидными плюсами, которые они несут, нужно не забывать о том, что при этом

увеличиваются риски, связанные с возможным распространением в сетях вредоносных компьютерных программ. В таких условиях понимание закономерностей распространение компьютерных вирусов, а также умение моделировать соответствующие процессы становится все более актуальным.

Распространение вируса в компьютерной сети подчиняется тем же законам, что и распространение инфекции в популяции живых организмов. Поэтому логично в качестве базовой (первоначальной) модели для описания процесса заражения вирусом компьютерной сети выбрать SI-модель - наиболее простую классическую модель распространения эпидемии [1] [2] [3]. В этой модели учитываются только наиболее значимые факторы, и делается много допущений, упрощающих модель.

На основе SI-модели построим иерархию пяти математических моделей, более корректно описывающих распространение компьютерного вируса. При этом, каждая следующая модель будет отличаться от предыдущей ведением в рассмотрение нового фактора.

При построении математической модели всегда возникает вопрос о том, насколько адекватно она отражает реальный ситуацию. Для подтверждения корректности модели необходимо сравнение результатов аналитического моделирования с данными, полученными в результате натурного эксперимента.

В качестве эмпирических данных для проверки адекватности создаваемой модели распространения вируса в компьютерной сети воспользуемся результатами наблюдений распространения в июле 2001 года интернет-червя Code-Red, которые опубликовали в 2002 году Дэвид Мур, Коллин Шеннон и Джеффри Браун [4]. Соответствующие статистические данные представлены на Рис. 1.

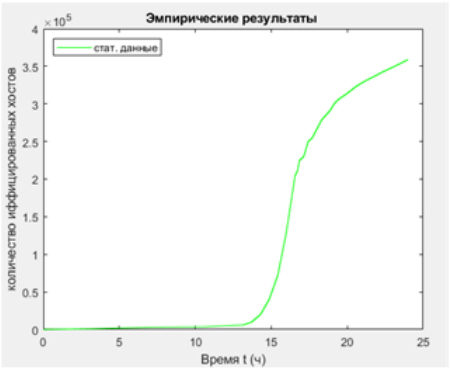

Рис. 1: *Общее количество уникальных IP-адресов, зараженных первой вспышкой Code-Red*

При построении иерархии математических моделей будем сопоставлять аналитические и эмпирические данные, а именно будем отслеживать, каким образом введение в рассмотрение того или иного параметра влияет на различие между результатами моделирования и экспериментальными данными, представленными на Рис. 1.

1. Модель №1

Сформулируем предположения, лежащие в основе модели №1:

– все хосты в сети делятся на два класса: восприимчивые к заражению и инфицированные;

– заражение происходит в результате «контакта» инфицированного хоста с восприимчивым;

– вероятность заражения при контакте постоянна;

– модель не предполагает возможность лечения зараженных хостов.

Введем следующие обозначения:

– N – общее количество хостов,

– $S(t)$ – количество восприимчивых хостов в момент времени t,

– $I(t)$ – количество инфицированных хостов в момент времени t.

Очевидно, что в любой момент времени должно выполняться соотношение:

$$S(t) + I(t) = N. \tag{1}$$

Будем считать, что вероятность контакта за единицу времени между двумя хостами равна $\frac{1}{N}$, а вероятность заражения при контакте – β. Тогда за время Δt одним инфицированным хостом будет заражено $\frac{\beta}{N} S(t)\Delta t$ компьютеров, значит, количество зараженных хостов в момент $t + \Delta t$ равно:

$$I(t + \Delta t) = I(t) + \frac{\beta}{N} S(t)I(t)\Delta t. \tag{2}$$

Следовательно,

$$\frac{I(t + \Delta t) - I(t)}{\Delta t} = \frac{\beta}{N} S(t)I(t). \tag{3}$$

Переходя к пределу при $\Delta t \to 0$, будем иметь:

$$\frac{dI}{dt} = \frac{\beta}{N} S(t)I(t). \tag{4}$$

Аналогично

$$\frac{dS}{dt} = -\frac{\beta}{N} S(t)I(t). \tag{5}$$

Тогда процесс распространения компьютерного вируса будет описываться следующей системой дифференциальных уравнений,

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta}{N} S(t)I(t), \\ \frac{dI}{dt} = \frac{\beta}{N} S(t)I(t). \end{cases} \tag{6}$$

Сравним результаты численных расчетов, полученные на основе данной модели, с имеющимися статистическими данными распространения компьютерного червя Code-Red в июле 2001 года.

Полагаем, что $S(0) = 362000$, $I(0) = 1$, $t_{start} = 0$, $t_{finish} = 24$. Численно подберем коэффициент β так, чтобы отличие между количеством зараженных хостов, полученным в результате математического моделирования, и статистическими данными было наименьшим. Таким коэффициентом является $\beta = 0.00000215$. Соответствующие графики количества зараженных хостов изображены на Рис. 2.

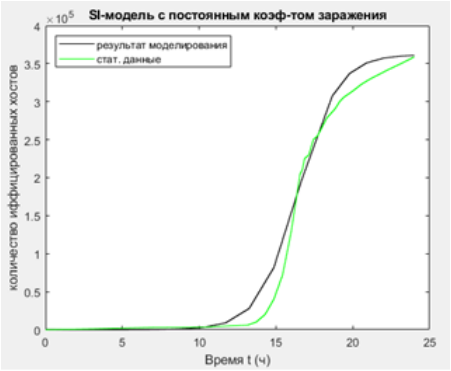

Рис. 2: *Сравнение количества зараженных хостов, полученного на основе SI-модели, с эмпирическими данными*

2. Модель №2

Заметим, что вероятность заражения восприимчивого хоста при контакте с инфицированным с течением времени уменьшается. Это связано с тем, что хосты пользователей, не проявляющих должной осторожности при работе в сети, заражаются в первые часы распространения вируса. То есть постепенно доля «беспечных» среди пользователей становится все меньше, и соответственно уменьшается вероятность заражения в ходе контакта. Для учета данного фактора будем предполагать, что коэффициент β - слабо убывающая функция, зависящая от времени.

Сформулируем предположения, лежащие в основе модели №2:

- все хосты в сети делятся на два класса: восприимчивые к заражению и инфицированные;
- заражение происходит в результате «контакта» инфицированного хоста с восприимчивым;
- вероятность заражения при контакте - монотонно убывающая функция;
- модель не предполагает возможность лечения зараженных хостов.

Полагаем, как и ранее, что $S(0) = 362000$, $I(0) = 1$, $t_{start} = 0$, $t_{finish} = 24$. Положим, $\beta = 0.00000215(1-0.0015t)$. Количество зараженных хостов, полученное в результате численных расчетов на основе модели №2, а также эмпирические данные изображены на Рис. 3.

Очевидно, что модель №2 дает результат более согласующийся с эмпирическими данными.

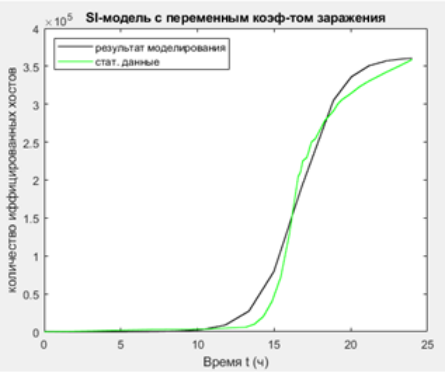

Рис. 3: *Сравнение результатов, полученных на основе SI-модели с переменным коэффициентом заражения, с эмпирическими данными*

3. Модель №3

В рассмотренных выше моделях не учитывается тот факт, что хост не сможет инфицировать другие компьютеры, пока не будет заражен окончательно. В связи с этим нужно разделить класс инфицированных на две группы: заразные (способны заразить восприимчивый хост) и латентные (инфицированные, но не заразные).

Сформулируем предположения, лежащие в основе модели №3:

– все хосты в сети делятся на три класса: восприимчивые, латентные и заразные;

– восприимчивый хост в результате «контакта» с заразным и последующим инфицированием сначала становится латентным, а через определенное время заразным;

– время полного заражения хоста постоянно;

– вероятность инфицирования при контакте - монотонно убывающая функция;

– модель не предполагает возможность лечения зараженных хостов.

Введем следующие обозначения:

– N – общее количество хостов,

– $S(t)$ – количество восприимчивых хостов в момент времени t,

– $I(t)$ – количество заразных хостов в момент времени t,

– $L(t)$ – количество латентных хостов в момент времени t,

– h – время полного заражения.

Модель, основанная на перечисленных выше предположениях, описывается следующей системой:

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta(t)}{N} S(t)I(t), \\ \frac{dL}{dt} = \frac{\beta(t)}{N} S(t)I(t), \\ I(t) = L(t-h). \end{cases} \tag{7}$$

которая сводится к системе дифференциальных уравнений с постоянным запаздыванием

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta(t)}{N} S(t)L(t-h), \\ \frac{dL}{dt} = \frac{\beta(t)}{N} S(t)L(t-h). \end{cases} \tag{8}$$

На Рис. 4 представлены используемые ранее статистические данные и результаты численных расчетов (а именно, количество зараженных хостов), полученные на основе модели №3 в случае $h = 0,03$ (остальные коэффициенты оставлены без изменения).

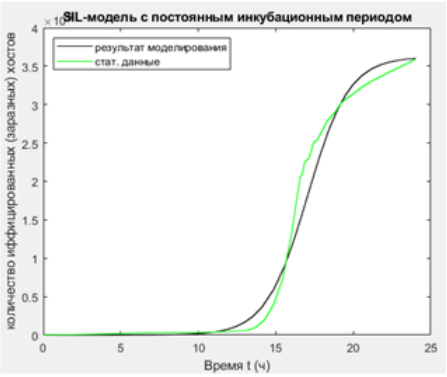

Рис. 4: Сравнение результатов, полученных на основе SIL-модели с постоянным инкубационным периодом, с эмпирическими данными

Отметим, что учет латентного периода меняет динамику распространения вируса (увеличивается время распространения вируса), а также, что модель №3 в большей степени соответствует эмпирическим данным, по сравнению с моделью №2, не учитывающей период латентности.

4. Модель №4

Компьютерный вирус – «живая» конструкция, он не остается неизменным, а совершенствуется в процессе своего существования, поэтому с течением времени процесс заражения инфицированного хоста происходит быстрее, то есть инкубационный период уменьшается. Следовательно, время полного заражения логичнее считать не постоянной величиной h, а слабо убывающей функцией времени $h(t)$, а соответствующая математическая модель будет иметь вид:

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta(t)}{N} S(t)L(t-h(t)), \\ \frac{dL}{dt} = \frac{\beta(t)}{N} S(t)L(t-h(t)). \end{cases} \tag{9}$$

Пусть в условиях предыдущей задачи $h(t) = 0.03 - 0.0008t$. Соответствующий график количества зараженных хостов представлен на Рис. 5.

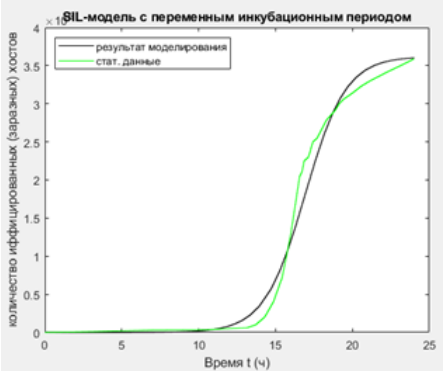

Рис. 5: *Сравнение результатов, полученных на основе SIL-модели с переменным инкубационным периодом, с эмпирическими данными*

График, соответствующий модели №4 еще более точно описывает процесс распространения компьютерного вируса (разница между графиком количества зараженных хостов и эмпирическими данными уменьшилась).

Динамику изменения количества зараженных компьютеров при постепенном переходе от модели №1 к модели №4 можно проследить по Рис. 6.

5. Модель №5

Модель распространения компьютерного вируса в сети без учета использования антивирусных баз, конечно же, является неполной. Поэтому следующим шагом в построении адекватной модели является введение в нее возможности лечения зараженных хостов.

Отметим, что антивирусное программное обеспечение может излечить хост только от тех вирусов, сигнатуры которых внесены в его базу. Будем считать, что и время обнаружения известного вируса путем сканирования и его лечение пренебрежимо малы. То есть ликвидация вируса происходит практически сразу, а значит, распространения инфекции не происходит. Поэтому динамику обнаружения и лечения известных вирусов рассматривать нет смысла.

Если же данный вирус не определяется антивирусным программным обеспечением, то происходит его распространение по сети. Эпидемия нарастает, пока информация о новом вирусе не попадет в антивирусные базы при их очередном обновлении.

Отметим среди существующих математических моделей распространения компьютерных вирусов модель PSIDR [5], которая описывает распространение вредоносного ПО и выделяет два временных периода распространения вируса.

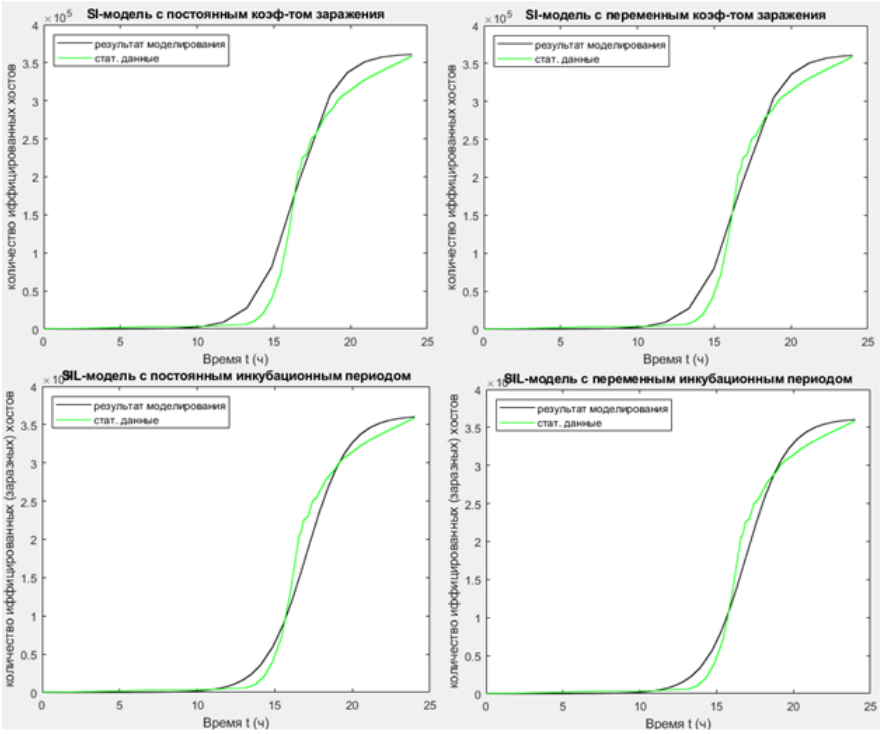

Рис. 6: *Сравнение результатов, полученных на основе моделей №№1-4.*

Построим новую модель на основе описанной выше модели №4, используя некоторые элементы PSIDR-модели.

Сформулируем предположения, лежащие в основе модели №5:

– все хосты в сети делятся пять классов: восприимчивые, латентные, заразные, находящиеся на излечении и получившие иммунитет после излечения;

– модель рассматривает два периода распространения вируса: предварительный период – от момента попадания вируса в сеть до момента, когда происходит обнаружение вируса антивирусным программным обеспечением, и период отклика - с момента обнаружения вируса антивирусным ПО

– в предварительный период вирус распространяется согласно модели №4;

– в период отклика после обнаружения вируса и занесения его в базы антивирусных средств начинается постепенное излечивание зараженных хостов, а незаражённые хосты приобретают невосприимчивость к вирусу, данный процесс характеризуется определенной скоростью, зависящей от скорости сканирования данного антивирусного ПО, а также временем необходимым на излечивание каждого хоста.

Введем следующие обозначения:

– $S(t)$ – количество восприимчивых хостов в момент времени t,

– $I(t)$ – количество заразных хостов в момент времени t,

– $L(t)$ – количество латентных хостов в момент времени t,

– $D(t)$ – количество хостов, находящихся на излечении в момент времени t,

– $R(t)$ – количество хостов, получивших иммунитет в момент времени t,

– $\beta(t)$ – вероятность заражения при контакте,

– $h(t)$ – время полного заражения,

– τ – момент обнаружения вируса (занесения его в базы данных антивирусных средств),

– m – коэффициент, характеризующий скорость сканирования сети данным антивирусным ПО,

– p – время полного излечения.

Будем считать, что за время Δt будет сканировано $m(S(t) + I(t))\Delta t$ хостов из числа зараженных или восприимчивых, значит, количество хостов, находящихся на излечении в момент $t + \Delta t$ равно:

$$D(t + \Delta t) = D(t) + m(S(t) + I(t))\Delta t. \tag{10}$$

Следовательно,

$$\frac{dD}{dt} = m(S(t) + I(t)). \tag{11}$$

Через время p каждый хост, находящийся на излечении, переходит в разряд «получивших иммунитет», то есть;

$$R(t) = D(t - p), \tag{12}$$

а значит,

$$\frac{dR}{dt} = m(S(t - p) + I(t - p)) \tag{13}$$

или

$$\frac{dR}{dt} = m(S(t - p) + L(t - p - h(t))). \tag{14}$$

Следовательно, имеем следующую дифференциальную систему, описывающую математическую модель:

предварительный период $t \leq \tau$:

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta(t)}{N} S(t)L(t - h(t)), \\ \frac{dL}{dt} = \frac{\beta(t)}{N} S(t)L(t - h(t)). \end{cases} \tag{15}$$

период отклика $t > \tau$:

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta(t)}{N} S(t)L(t - h(t)) - mS(t - p), \\ \frac{dL}{dt} = \frac{\beta(t)}{N} S(t)L(t - h(t)) - mL(t - p - h(t)) \\ \frac{dR}{dt} = m(S(t - p) + L(t - p - h(t))). \end{cases} \tag{16}$$

Фиксируем следующие значения коэффициентов: $n = 500$, $S(0) = 495$, $R(0) = 0$, $I(0) = 5$, $\tau = 40$, $p = 0.00001$, $m = 0.04$, $h(t) = 0$, $\beta(t) = 0.00022(1 - 0.00015t)$.

На Рис. 7 и Рис. 8 представлены соответствующие результаты численного эксперимента на основе построенной модели №5. На Рис. 7 – динамика количества зараженных хостов (численность группы, отслеживаемой во всех предыдущих моделях). На Рис. 8 – динамика численности всех групп, участвующих в эпидемиологическом процессе.

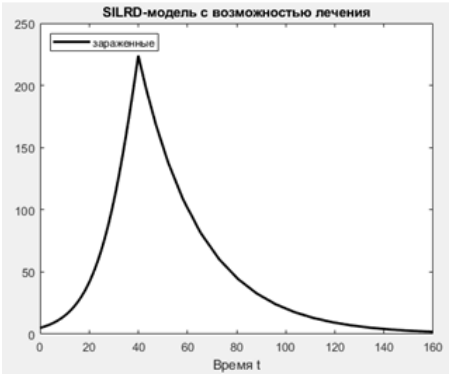

Рис. 7: *Количество зараженных хостов, рассчитанное на основе SILDR-модели*

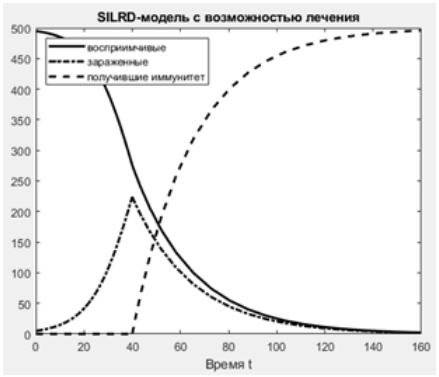

Рис. 8: *Численности групп, рассматриваемых в SILDR-модели*

Заключение

В статье построена иерархия из пяти математических моделей, описывающих распространение компьютерного вируса в сети. Первые четыре модели не предполагают возможность лечения зараженных хостов и опираются на SI-модель распространения эпидемии, при этом, каждая следующая модель отличается от предыдущей введением в рассмотрение нового фактора.

Для каждой из этих моделей проведено сравнение результатов численного моделирования с экспериментальными данными, опубликованными в 2002 году Д.

Муром, К. Шенноном и Д. Брауном. При этом продемонстрировано, что при постепенном переходе от модели №1 к №4 уменьшаются различия между результатами моделирования и экспериментальными данными, что подтверждает рост корректности.

Последняя из представленных в статье моделей (модель №5) описывает распространение вируса в компьютерной сети с учетом работы антивирусного программного обеспечения. Модель сочетает в себе элементы модели №4 и известной PSIDR-модели. Модель № 5 рассматривает пять классов, участвующих в эпидемиологическом процессе (восприимчивые, латентные, заразные, находящиеся на излечении, получившие иммунитет после излечения) и два периода распространения вируса (первый – от момента попадания вируса в сеть до момента обнаружения вируса антивирусным ПО и второй – с момента обнаружения вируса и занесения его в базы антивирусных средств).

В работе представлены результаты численного эксперимента, позволяющие проследить динамику численности всех групп, участвующих в эпидемиологическом процессе, описываемом моделью №5.

Список литературы

[1] Kermack W.O., McKendrick A.G. A Contribution to the Mathematical Theory of Epidemics // Proceedings of the Royal Society. 1927. Vol. 115, № 772. Pp. 700–721.

[2] Гусаров А.Н., Жуков Д.О., Косарева А.В. Описание динамики распространения компьютерных угроз в информационно- вычислительных сетях с запаздыванием действия антивирусов // Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия Приборостроение. 2010. № 1 (78). С. 112–120.

[3] Семыкина Н.А. Математическое моделирование распространения вирусов с учетом влияния временных параметров // Перспективы развития информационных технологий. 2013. № 14. С. 139–144.

[4] Moore D., Shannon C., Brown J. Code-Red: a case study on thespread and victims of an Internet Worm // Proc. of ACM/USENIX Internet Measurement Workshop. 2002.

[5] Минаев В.А., Вайц Е.В., Корячко А.В., Киракосян А.Э. Системно-динамическое моделирование распространения компьютерных вирусов // Технологии техносферной безопасности. 2017. № 3 (73). С. 220–229.

Образец цитирования

Еремеева Н.И. Построение модели распространения вируса в компьютерной сети на основе сравнения результатов моделирования с эмпирическими данными // Вестник ТвГУ. Серия: Прикладная математика. 2022. № 4. С. 39–52. https://doi.org/10.26456/vtpmk647

Сведения об авторах

1. **Еремеева Нина Игоревна**
доцент кафедры высшей математики Димитровградского инженерно-технологического института – филиала федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ».

Россия, 433511, Ульяновская обл., г. Димитровград, ул. Куйбышева, д. 294.
E-mail: vm-diti.mifi@yandex.ru

BUILDING A VIRUS PROPAGATION MODELIN A COMPUTER NETWORK BASED ON A COMPARISON OF RESULTS SIMULATIONS WITH EMPIRICAL DATA

Eremeeva Nina Igorevna
Associate Professor at the Department of
Dimitrovgrad Engineering and Technological Institute,
National Research Nuclear University MEPhI
Russia, 433511, Ulyanovsk region, Dimitrovgrad, Kuibyshev str., 294.
E-mail: vm-diti.mifi@yandex.ru

Received 21.11.2022, revised 20.12.2022.

The article uses the SI model to construct a hierarchy of four mathematical models describing the spread of a computer virus on the network in the absence of the possibility of treating infected hosts. Each subsequent model differs from the previous one by introducing a new factor into consideration. For the constructed models, numerical simulation results were compared with experimental data (statistical data published in 2002 by D. Moore, K. Shannon and D. Brown). Based on the constructed models using some elements of the PSIDR model, a model was created describing the spread of a computer virus in the network if there is a possibility of treating infected hosts, and a numerical experiment was carried out to trace the dynamics of all groups involved in the epidemiological process.

Keywords: mathematical modeling, systems of differential equations, virus propagation in a computer network.

Citation

Eremeeva N.I., "Building a virus propagation modelin a computer network based on a comparison of results simulations with empirical data", *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2022, № 4, 39–52 (in Russian). https://doi.org/10.26456/vtpmk647

References

[1] Kermack W.O., McKendrick A.G., "A Contribution to the Mathematical Theory of Epidemics", *Proceedings of the Royal Society*, **115**:772 (1927), 700–721.

[2] Gusarov A.N., Zhukov D.O., Kosareva A.V., "Description of the dynamics of the spread of computer threats in information and computing networks with a delay in the action of antiviruses", *Vestnik Moskovskogo gosudarstvennogo tekhnicheskogo universiteta im. N.E. Baumana. Seriya Priborostroenie [Bulletin of the Bauman Moscow State Technical University. Instrumentation Series]*, 2010, № 1 (78), 112–120 (in Russian).

[3] Semykina N.A., “Mathematical modeling of virus spread taking into account the influence of time parameters”, *Perspektivy razvitiya informatsionnykh tekhnologij [Prospects for the development of information technology]*, 2013, № 14, 139–144 (in Russian).

[4] Moore D., Shannon C., Brown J., “Code-Red: a case study on thespread and victims of an Internet Worm”, *Proc. of ACM/USENIX Internet Measurement Workshop*, 2002.

[5] Minaev V.A., Vajts E.V., Koryachko A.V., Kirakosyan A.E., “System-dynamic simulation of the spread of computer viruses”, *Tekhnologii tekhnosfernoj bezopasnosti [Technosphere security technologies]*, 2017, № 3 (73), 220–229 (in Russian).