

УДК 004.056.57

© 2011 г. **Я.М. Далингер**, канд. тех. наук

(Санкт-Петербургский государственный университет гражданской авиации),

Д.В. Бабанин

(Московский государственный институт электроники и математики),

С.М. Бурков, д-р техн. наук

(Тихоокеанский государственный университет, Хабаровск)

МАТЕМАТИЧЕСКИЕ МОДЕЛИ РАСПРОСТРАНЕНИЯ ВИРУСОВ В КОМПЬЮТЕРНЫХ СЕТЯХ РАЗЛИЧНОЙ СТРУКТУРЫ

В статье приведены описание и анализ математических моделей распространения компьютерных вирусов. Модели построены на основе цепей Маркова. Модели позволяют исследовать процесс распространения эпидемии компьютерных вирусов в сетях с различной структурой. Результаты могут быть полезны администраторам компьютерных сетей.

Ключевые слова: вредоносное программное обеспечение, защита данных, марковский процесс, компьютерная сеть, компьютерные вирусы.

Введение

Проблемы информационной безопасности требуют внимания не только во время эксплуатации сетей, но и на этапе их проектирования. Конечно, компьютерная вирусология обязана своим происхождением прежде всего организации операционных систем, разработанных корпорацией Микрософт, однако, учитывая распространение этих систем, анализ и решение проблем, возникающих в процессе построения компьютерных сетей, в настоящий момент являются актуальной задачей. Для оценки защищенности компьютерной сети от программ нарушителей необходимо иметь модель их распространения в сетях различной структуры. На данный момент существуют модели распространения вирусов в компьютерных сетях как заимствованные из эпидемиологии (SI, SIR [1]), так и разработанные с учетом особенностей компьютерных вирусов (AAWP [2], PSIDR [3]). Но, к сожалению, данные модели не позволяют учитывать структуру сети, так как предполагается, что структура сети описывается полносвязным графом. В данной работе ставится задача разработать модель распространения вредоносного программного обеспечения одного типа по сетям с различной архитектурой, без защиты на основе цепей Маркова. Моделирование параметров инфицирования, учет множества конкретных моделей программ-нарушителей позволят разработать оптимальную комплексную модель защиты данных для различных вычислительных систем.

Построение моделей

Рассмотрим локальную сеть, состоящую из N компьютеров. Каждый компьютер может находиться в одном из двух состояний – незараженный или зараженный.

Сеть можно представить в виде графа, узлами которого являются компьютеры, а дугами – каналы связи между ними, по которым могут распространяться вирусы. Вес связи w_{ij} означает вероятность перехода вируса по каналу связи между компьютерами i и j за единицу времени.

Модель на основе цепи Маркова для всей сети. Общее состояние сети в момент времени t является совокупностью состояний всех узлов сети. Оно может быть описано вектором из N элементов, где значение k -го элемента соответствует состоянию k -го узла: **I** (infected), если он заражен, и **S** (suspected), если не заражен.

Состояние сети в следующий момент времени зависит только от текущего состояния сети и не зависит от предыдущих. Следовательно, процесс распространения вируса в сети можно представить в виде цепи Маркова [4].

Переходные вероятности вычисляются по формуле:

$$P_{ij} = P[f^t = s^j \mid f^{t-1} = s^i], \quad (1)$$

Сеть перейдет из состояния s^i в состояние s^j при условии, если каждый узел сети перейдет из состояния s_k^i в состояние s_k^j , где k – номер узла. Вероятность этого события описывается следующей формулой:

$$\begin{aligned} P_{ij} &= P[f^t = s^j \mid f^{t-1} = s^i] = \\ &= P[f_1^t = s_1^j \cap f_2^t = s_2^j \cap \dots \cap f_N^t = s_N^j \mid f_1^{t-1} = s_1^i \cap f_2^{t-1} = s_2^i \cap \dots \cap f_N^{t-1} = s_N^i] = (2) \\ &= \prod_{k=1}^N P[f_k^t = s_k^j \mid f_k^{t-1} = s_k^i]. \end{aligned}$$

Для определения вероятности

$$P[f_k^t = s_k^j \mid f_k^{t-1} = s_k^i]$$

перехода k -го компьютера из состояния s_k^i в состояние s_k^j необходимо рассмотреть четыре варианта для различных состояний компьютера на предыдущем и текущем шаге: переход из состояния **S** в состояние **I**, из **S** \rightarrow **S**, **I** \rightarrow **S** и **I** \rightarrow **I**.

S \rightarrow **I**. Вероятность заражения незараженного k -го компьютера при исходном состоянии сети s^i равна $P_{зар}(k, s^i)$.

S \rightarrow **S**. Вероятность того, что компьютер останется незараженным будет равна $[1 - P_{зар}(k, s^i)]$, поскольку события перехода компьютера из **S** в **I** и перехода из **S** в **S** образуют полную группу событий.

I \rightarrow **S**. Вероятность излечения компьютера равна нулю, так как модель не учитывает лечения.

I \rightarrow **I**. Вероятность того, что компьютер останется зараженным, равна единице, так как модель не учитывает «лечения».

В результате получена следующая формула:

$$P[f_k^t = s_k^j | f_k^{t-1} = s_k^i] = \begin{cases} P_{\text{зап}}(k, s^i), & \text{если } s_k^i = S, s_k^j = I, \\ 1 - P_{\text{зап}}(k, s^i), & \text{если } s_k^i = S, s_k^j = S, \\ 0, & \text{если } s_k^i = I, s_k^j = S, \\ 1, & \text{если } s_k^i = I, s_k^j = I. \end{cases} \quad (3)$$

Вероятность передачи вируса от узла m узлу k при состоянии сети s_m^i можно вычислить следующим образом:

если компьютер m заражен, вероятность равна w_{mk} (как это было определено в начале статьи);

если компьютер m не заражен, то вероятность передачи вируса с него равна нулю.

$$P_{\text{передачи}}(m, k, s_m^i) = \begin{cases} w_{mk}, & \text{если } s_m^i = I, \\ 0, & \text{если } s_m^i = S. \end{cases} \quad (4)$$

Узел k перейдет из незараженного состояния в зараженное за единицу времени в том случае, если вирус к нему поступит хотя бы с одного другого узла. Поскольку события заражения k -го узла от различных источников являются независимыми, то вероятность заражения незараженного k -го узла будет равна:

$$P_{\text{зап}}(k, s^i) = 1 - \prod_{m=1}^N (1 - P_{\text{передачи}}(m, k, s_m^i)). \quad (5)$$

Модель на основе цепи Маркова для отдельных узлов. Если рассматривать в виде марковской цепи не процесс распространения вируса по всей сети, а строить отдельную марковскую цепь для каждого узла, можно значительно сократить объем вычислений.

В каждый момент времени каждый компьютер с определенной вероятностью может быть незараженным (S), либо зараженным (I). Вектор состояния в данном случае состоит из двух элементов – вероятности того, что компьютер не заражен, и вероятности того, что компьютер заражен: $p_k^t = \{P_k^t(S), P_k^t(I)\}$, где k – номер компьютера, t – номер шага.

Матрица переходных вероятностей для данного узла будет иметь следующий вид:

$$P = \begin{pmatrix} P(f^t = S | f^{t-1} = S) & P(f^t = I | f^{t-1} = S) \\ P(f^t = S | f^{t-1} = I) & P(f^t = I | f^{t-1} = I) \end{pmatrix}. \quad (6)$$

Поскольку данная модель не учитывает возможности излечения, то переход из состояния I в состояние S невозможен, а из состояния I можно попасть только обратно в состояние I, то получим:

$$P[f^t = S | f^{t-1} = I] = 0, \quad P[f^t = I | f^{t-1} = I] = 1.$$

Так как сумма элементов строки матрицы переходов всегда равна единице, то имеем:

$$P[f^t = S | f^{t-1} = S] = 1 - P[f^t = I | f^{t-1} = S].$$

Обозначив $P_{\text{зап}}(k) = P[f^t = I | f^{t-1} = S]$, где k – это номер узла, для которого составляется модель, получаем следующий вид матрицы переходов:

$$\mathbf{P} = \begin{pmatrix} 1 - P_{\text{зар}}(k) & P_{\text{зар}}(k) \\ 0 & 1 \end{pmatrix} \quad (7)$$

где $P_{\text{зар}}(k)$ – это вероятность заражения k -го узла, которая, как было показано выше, вычисляется по формуле

$$P_{\text{зар}}(k) = 1 - \prod_{m=1}^N (1 - P_{\text{передачи}}(m, k)). \quad (8)$$

Передача вируса от узла m узлу k произойдет при одновременном наступлении следующих событий:

если компьютер m заражен на предыдущем шаге, вероятность этого события равна $P_k^{t-1}(I)$;

если вирус пройдет по связи $m-k$, вероятность этого события равна w_{mk} (как это было определено в начале статьи).

Следовательно, вероятность передачи вируса от узла m узлу k равна произведению вероятности заражения компьютера m на предыдущем шаге, умноженной на вероятность перехода вируса по связи $m - k$:

$$P_{\text{передачи}}(m, k) = P_m^{t-1}(I) w_{mk}. \quad (9)$$

В результате подстановки получаем следующую матрицу переходов для отдельного узла:

$$\mathbf{P} = \begin{pmatrix} \prod_{m=1}^N (1 - P_m^{t-1}(I) \cdot w_{mk}) & 1 - \prod_{m=1}^N (1 - P_m^{t-1}(I) \cdot w_{mk}) \\ 0 & 1 \end{pmatrix}. \quad (10)$$

Как можно заметить, матрица переходов зависит от состояния узлов сети на предыдущем шаге, следовательно, цепь Маркова для каждого узла является неоднородной.

Для использования модели на основе цепи Маркова необходимо задать начальное распределение π_0 – вектор вероятностей нахождения сети в том или ином состоянии в начальный момент времени. Выбирается единственное начальное состояние сети f_0 , для которого вероятность принимается равной единице, для остальных – нулю:

$$p_0 = \{p_j^0\}, \quad p_j^0 = P[f_0 = s_j] = \begin{cases} 1, & f_0 = s_j \\ 0, & f_0 \neq s_j \end{cases}.$$

Исходя из теории марковских цепей, распределение на шаге t будет равно $p_t = p_{t-1} \mathbf{P}$.

Математическое ожидание среднего числа зараженных компьютеров на шаге n можно вычислить следующим образом.

Для каждого состояния сети s_j легко определить количество зараженных компьютеров: поскольку s_j представляет собой вектор, состоящий из N элементов, количество зараженных компьютеров для состояния s_j определяется как

$$N_I(s_j) = \sum_{i=1}^N \begin{cases} 1, & s_{ij} = I, \\ 0, & s_{ij} \neq I. \end{cases}$$

Поскольку сумма всех элементов вектора состояния на шаге t всегда равна

единице, то математическое ожидание количества зараженных компьютеров будет равно

$$M[N_I^t] = \sum_{j=1}^{2^N} p_j^{(t)} \cdot N_I(s_j).$$

Проводя вычисления для $t = 0 \dots t_{max}$, получим зависимость среднего количества зараженных компьютеров от номера шага t .

Использование модели для отдельных узлов значительно проще. Начальное распределение задается для каждого узла сети: $p_j^0 = \{P_j^0(S), P_j^0(I)\}$, где j – номер узла. Наиболее удобный способ построения начальных распределений – выбрать в сети зараженные компьютеры, для которых вероятность нахождения в зараженном состоянии равна 1, т.е. $p_j^0 = \{0; 1\}$, а для остальных – наоборот, $p_j^0 = \{1; 0\}$.

Далее для каждого шага t производятся следующие действия:

для каждого j -го узла по формуле (10) строится матрица переходов;

вектор начального на предыдущем шаге умножается на полученную матрицу переходов, в результате получается вектор распределения для первого шага:

$$p_j^t = p_j^{t-1} \mathbf{P};$$

имея множество векторов распределения на шаге t $\{p_1^t, p_2^t, \dots, p_N^t\}$, а следовательно, зная вероятность нахождения каждого узла в зараженном состоянии в момент времени t ($P_j^t(I)$), можно вычислить математическое ожидание количества зараженных компьютеров на этом шаге:

$$M[N_I^t] = \sum_{j=1}^N P_j^t(I).$$

Проводя вычисления для $t = 0 \dots t_{max}$, получим зависимость среднего количества зараженных компьютеров от номера шага t .

Моделирование

Для моделирования с использованием данных моделей была разработана программа на языке Java, обладающая следующей функциональностью:

создание и редактирование структуры сети (ориентированного графа);

выбор изначально зараженных узлов;

использование различных моделей распространения вирусов (цепь Маркова для всей сети, цепи Маркова для отдельных узлов, SI-модель);

обработка результата, построение графиков по различным моделям.

В качестве входных данных моделирования используется следующая информация:

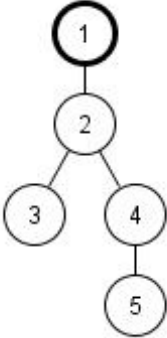
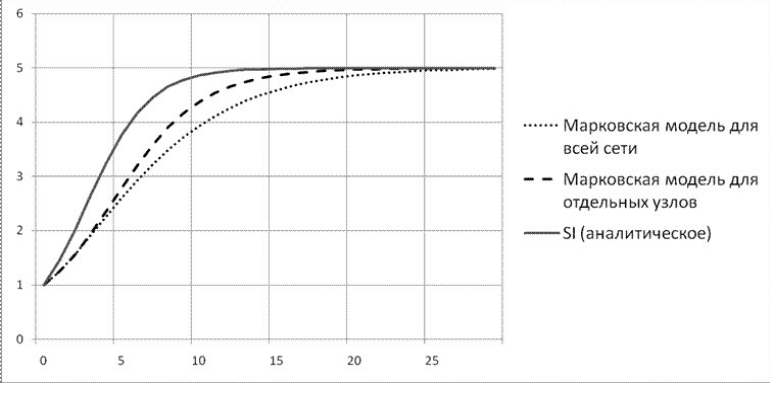
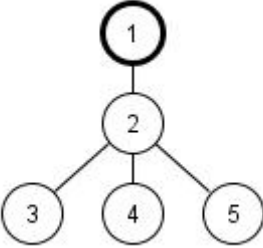
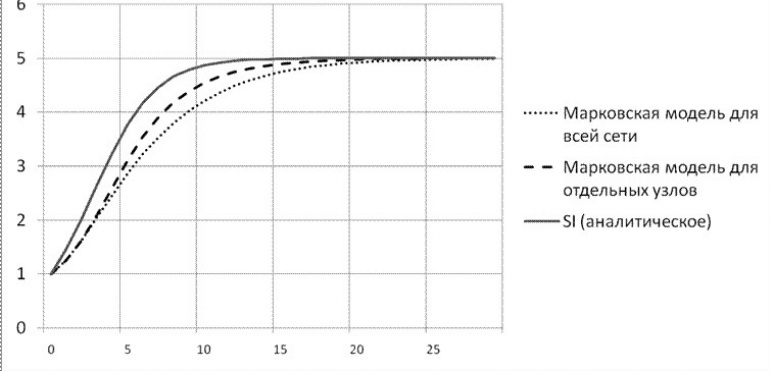
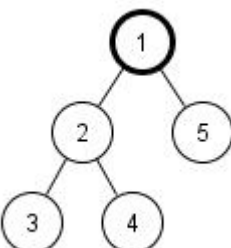
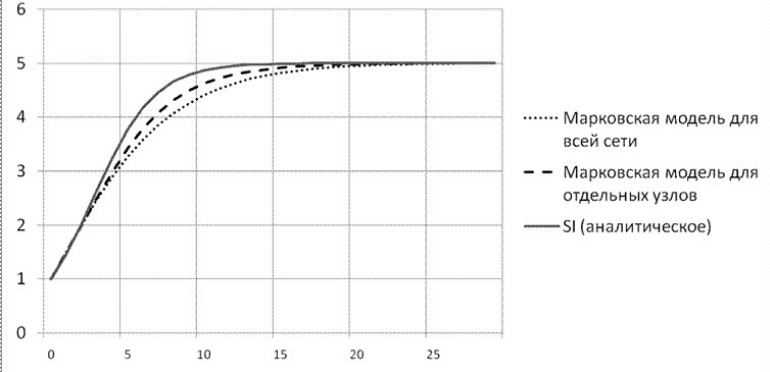
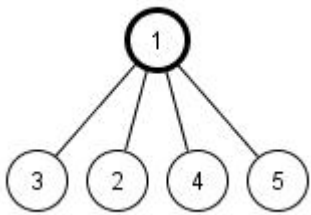
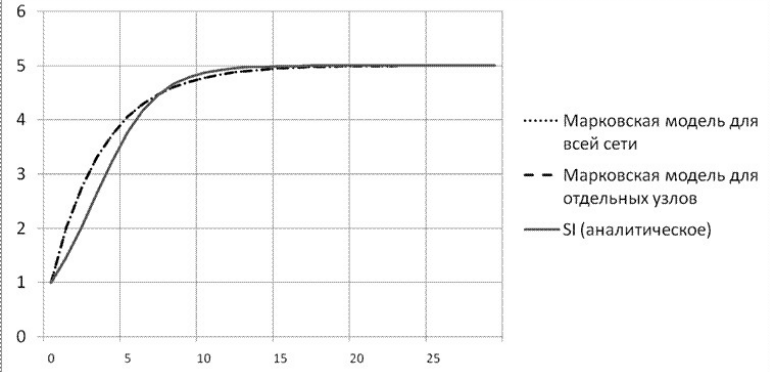
количество узлов в сети;

связи между узлами сети (информационные связи, по которым могут распространяться вирусы);

веса связей (вероятность передачи вируса с зараженного узла на незараженный за один шаг через данную связь);

начальное состояние сети (зараженные узлы).

Таблица 2

Структура сети	Зависимость количества зараженных узлов от номера шага
<p>№ 1</p> 	
<p>№ 2</p> 	
<p>№ 3</p> 	
<p>№ 4</p> 	

На графиках можно заметить, что при использовании модели SI результат одинаков вне зависимости от структуры сети. Это обусловлено тем, что модель SI учитывает только информацию о количестве узлов в сети и скорости заражения. Поскольку количество узлов, количество связей и весовые коэффициенты одина-

ковы, то и результаты моделирования для таких исходных данных совпадают.

Графики зависимостей, полученные на основе моделей, использующих цепи Маркова, изменяются в зависимости от структуры сети.

Для структуры 1 количество зараженных узлов увеличивается медленнее, чем для остальных. Это объясняется тем, что для того, чтобы заразить узел 4, вирус должен прежде заразить узлы 1 и 3, для чего необходимы, как минимум, два шага.

Структура 2 характеризуется более быстрым нарастанием количества инфицированных узлов. Это произошло потому, что наиболее удаленный от источника заражения узел 4 был перемещен ближе к источнику заражения.

Структура 3 показывает еще более быстрое распространение вируса по сети: узел 4 непосредственно связан с изначально зараженным узлом.

Худшей с точки зрения вирусной безопасности структурой сети с заданным количеством связей является структура 4. В ней каждый узел соединен с источником заражения, что обеспечивает очень быстрое распространение вируса. Это ясно видно на графике, где совпадающие кривые моделей на основе марковских цепей с начала моделирования идут гораздо более круто, чем графики SI-модели.

Достоинство данной модели – то, что модель оперирует однозначными состояниями всей сети. Благодаря этому в результате моделирования можно анализировать, какие сегменты сети будут заражены полностью, а какие – частично.

Использование однородной цепи Маркова дает возможность активно использовать математический аппарат теории марковских цепей. Возведя матрицу в определенную степень, можно рассчитать вероятности заражения узлов из любого начального состояния, умножив начальное распределение на полученную матрицу. Для различных начальных распределений можно использовать однажды построенную матрицу. Однако при любом изменении структуры сети построение матрицы переходов необходимо проводить заново.

Основным недостатком данной модели является очень большой объем вычислений. Для сети, состоящей из N узлов, существует 2^N различных состояний сети. Для расчетов по данной модели необходимо построить переходную матрицу. Ее размер будет составлять 2^{2N} элементов. Поскольку для каждого элемента матрицы при вычислении вероятности перехода происходит обход всех узлов сети, для каждого из которых также происходит обход всех узлов сети, сложность алгоритма построения матрицы переходных вероятностей составляет $O(N^2 2^{2N})$.

Недостатком, ограничивающим расширение данной модели, является резкое возрастание данных для вычислений при увеличении числа состояний каждого узла, т.е. при переходе от двух состояний – S и I – к модели с тремя состояниями (модель, учитывающая лечение – SIR) сложность алгоритма построения матрицы переходных вероятностей возрастает до $O(N^2 3^{2N})$.

Несомненное достоинство по сравнению с предыдущей моделью – гораздо меньший объем вычислений. Поскольку размер матрицы переходов для каждого узла составляет 2×2 , а на каждом шаге для каждого узла необходимо перебрать всех соседей, то сложность алгоритма составляет $O(N^2 T)$, где N – количество узлов в сети; T – количество шагов моделирования.

В отличие от модели, учитывающей состояние всей сети, данная модель оперирует лишь вероятностями заражения отдельных узлов. Поскольку модель обладает марковским свойством (т.е. для получения следующего состояния используется текущее и не используется информация о предыдущих состояниях), то, имея в виде результата моделирования вероятности заражения отдельных узлов, невозможно установить, какие компьютеры могут быть заражены одновременно, а какие нет.

Заключение

Предложенные модели позволяют рассчитывать распространение компьютерных вирусов в вычислительных сетях различной топологии. Был описан механизм их использования для получения информации о характере распространения вирусной эпидемии в сети. Было проведено моделирование заражения узлов для сетей различной структуры с использованием трех различных моделей, в результате чего получены зависимости количества инфицированных узлов от времени. Был проведен анализ результатов моделирования, а также сделано сравнение предложенных моделей между собой, определены их сильные и слабые стороны. Использование данных моделей позволит оценить защищенность от компьютерных вирусов сетей различных топологий, что позволит выбирать наиболее безопасные топологии сети на самом раннем этапе проектирования.

ЛИТЕРАТУРА

1. Jeffrey O. Kephart, Steve R. White. Directed-Graph Epidemiological Models of Computer Viruses // IEEE Symposium on Security and Privacy, 1991. – P. 343.
2. Zesheng Chen, Lixin Gao, and Kevin Kwait. Modeling the spread of active worms / Zesheng Chen, Lixin Gao, Kevin Kwait // INFOCOM 2003 [Электронный ресурс]. – 2003. – Режим доступа: http://www.ieee-infocom.org/2003/papers/46_03.PDF. – Дата доступа: 19.03.2010 г.
3. Williamson M.M., Leveille J. An epidemiological model of virus spread and cleanup / M. M. Williamson, J. Leveille // HPL-2003-39 [Электронный ресурс]. – 2003. – Режим доступа: <http://www.hpl.hp.com/techreports/2003/HPL-2003-39.pdf>. – Дата доступа: 19.03.2010 г.
4. Ревюз Д. Цепи Маркова. – М.: РФФИ, 1997.

Статья представлена к публикации членом редколлегии А.Д. Плутенко.

E-mail:

Далингер Яков Михайлович – iakovdalinger@gmail.com;

Бабанин Дмитрий Владимирович – saksmiem@mail.ru;

Бурков Сергей Михайлович – burkov@khh.ru.