УДК 004.738 (043.3)

Зиатдинов А.Ю.

МЕТОД ОБНАРУЖЕНИЯ АНОМАЛИЙ В СПЕЦИАЛИЗИРОВАННЫХ КОРПОРАТИВНЫХ СЕТЯХ

Институт компьютерных технологий
Национальный авиационный университет

Рассмотрена задача обнаружения аномалий в работе специализированной корпоративной сети банка или финансовой компании. Проведено сравнение методов детерминированного и статистического анализа отклонений от нормальных режимов работы сегментов сети. Предложены математические модели аномалий. Сделан вывод о необходимости сочетания детерминированных и статистических методов и предложен алгоритм последовательного обнаружения. Получены количественные оценки необходимых объемов выборок для достижения ошибок первого и второго рода не более заданных

Введение

Современные информационно-телекоммуникационные системы характеризуются большой сложностью и высокой стоимостью. Вследствие этого эвристические подходы к формированию архитектуры, выбору основных конструктивных и эксплуатационных характеристик и оценке параметров постепенно заменяются регулярными методами анализа и синтеза.

Если раньше при развертывании компьютерных сетей достаточно было пользоваться интуитивными соображениями и здравым смыслом, то теперь становится необходимым владеть математическим аппаратом, позволяющим рассчитать оптимальную структуру отдельных сегментов и устройств, а подчас и облик всей сети в целом. В сложных современных задачах при помощи интуиции и «физического смысла» удается сконструировать лишь весьма посредственные структуры, которые, как правило, могут быть заменены на более совершенные, если обратиться к систематической теории.

Необходимо применять, прежде всего, математические методы синтеза, расчета и проектирования с учетом специфики компьютерных сетей. С течением времени они приобретают все большее значение. Из теорий, помогающих в этих расчетах, следует упомянуть, в первую очередь, теорию массового обслуживания, теорию оптимального и адаптивного управления, линейной и нелинейной фильтрации, а также другие разделы математической статистики, теорию графов, сетей, теорию информации и др.

Другой важной проблемой является разработка основных математических методов и уравнений, удобных для решения конкретных практических сетевых задач.

Представление сети любого масштаба в виде детерминированной системы и описание её соответствующими уравнениями с детерминированными параметрами даст весьма грубый, практически бесполезный результат по следующим причинам.

Во-первых, необходимо иметь полную априорную информацию о текущих параметрах и состоянии сети. Такая задача является практически нереализуемой в подавляющем большинстве случаев.

Во-вторых, отказы оборудования, аномальные ситуации, нарушения в работе сети, перегрузки из-за перепадов сетевой и вычислительной нагрузки по определению являются случайными событиями, которые мы не можем контролировать и которыми невозможно управлять – их можно только прогнозировать с некоторой точностью.

В-третьих, даже в идеальном случае наличия полной априорной информации о параметрах, структуре и мгновенном состоянии сети эти данные будут практически бесполезны. Системы уравнений, ко-

торыми описывается сеть, будут иметь порядок, сравнимый с числом сетевых и терминальных узлов. Для численного решения такой системы уравнений в реальном масштабе времени потребуется практически недостижимый объем вычислительных ресурсов. Кроме того, можно утверждать, что ошибки расчетов будут расти до недопустимых величин, и полученный результат будет совершенно бесполезен.

Поэтому в настоящее время только статистические методы описания сетей, процессов обмена данными, синтеза структуры сети и оценки параметров, управления сетями могут давать результаты удовлетворительной точности. При правильном выборе достаточных статистик и методов их анализа потребные вычислительные ресурсы также оказываются вполне приемлемыми.

Обнаружение аномалий в работе информационно-вычислительных систем

При разработке систем обнаружения угроз работе компьютеров и компьютерных сетей исторически применялись детерминированные методы анализа. Пока число потенциальных угроз было сравнительно небольшим, детерминистский подход себя оправдывал. Например, в наиболее совершенных антивирусных программах анализ известных масок вирусов выполняется практически в реальном масштабе времени. Однако с течением времени число вирусов и других вредоносных программ (далее – вирусов) растет в геометрической прогрессии. Для обеспечения работы антивирусных программ, анализа файлов и мониторинга компьютера в реальном или хотя бы в квазиреальном масштабе времени, когда пользователь не ощущает задержек в работе компьютера или сети в целом, требуется все больший объем вычислительных ресурсов. Рано или поздно эти требования станут неприемлемыми, поскольку теоретически верхнего предела числу вирусов не существует.

Еще более серьезным недостатком детерминистского подхода является то, что до занесения маски нового вируса в базу данных антивирусной программы компьютер остается уязвимым. Чаще всего базы данных обновляются уже после того, как более или менее значительной группе пользователей нанесен ущерб. Иногда этот ущерб является огромным, а иногда – вообще неприемлемым (например, для систем обороны, энергетики и им подобных). По существу, защита всегда на шаг или даже на несколько шагов отстает от активности источников угроз, а обнаружение и предотвращение вновь возникающих угроз идет методом «проб и ошибок».

Такая же ситуация складывается и в сфере защиты от несанкционированного доступа (атак, вторжений). Основой традиционных систем обнаружения вторжений являются методы анализа сигнатур и/или анализа протоколов. И тот, и другой методы имеют свои достоинства и недостатки. Как правило, используются комбинированные системы обнаружения вторжений с совместным анализом сигнатур и протоколов, созданием правил обработки специфического (аномального, подозрительного) трафика и т.д. Однако и в этом случае защита находится в режиме пассивного ожидания новых угроз. После выявления этих угроз, также зачастую методом проб и ошибок, модернизируются базы данных сигнатур (протоколов) для принятия адекватных мер защиты. И так же, как в антивирусных программах, из-за неограниченного роста объема этих баз данных все большая часть вычислительных ресурсов уходит на нужды защиты. Таким образом, эффективность детерминированных систем обнаружения вторжений с течением времени постоянно снижается.

Логическим выходом из сложившейся ситуации является переход к статистическим методам анализа угроз и статистическому синтезу систем защиты.

Не являются исключением и корпоративные сети банков, инвестиционных

компаний, различных фондов и др. Ядром любой такой структуры является так называемый процессинговый центр, в котором решаются задачи хранения, обработки и передачи данных. Подавляющую часть этих данных составляют данные критического характера, утечка которых представляет непосредственную угрозу финансовой безопасности как клиентов, так и банка в целом.

Рассмотрим гипотетическую структуру сети крупного банка или финансовой компании (рис. 1). Основу ее составляет корпоративная сеть (КрС), связанная по различным линиям передачи данных (ЛПД) с терминальными узлами (ТУ). К терминальным узлам можно отнести автономные (по территориальному расположению) отделения и филиалы, а также множество банкоматов и платежных терминалов. Связь ТУ с КрС осуществляется провайдерами телекоммуникаций (ТЛК) и провайдерами Интернет.

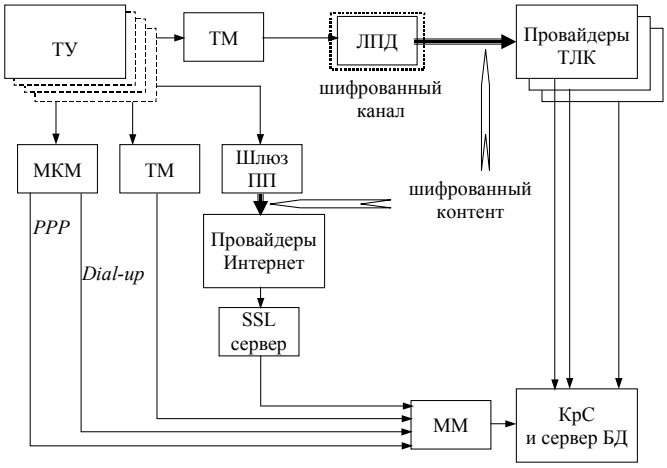

Рис. 1. Структура сети

ТМ – терминальный маршрутизатор;
МКМ – многоканальный модем;
ММ – магистральный маршрутизатор;
PPP, *Dial-Up* – протоколы передачи;
БД – база данных; шлюз ПП – шлюз пакетной передачи данных по протоколам *GPRS*/*EDGE*/3*G*.

Вероятностная модель процесса обнаружения сетевых аномалий (СА) в сети имеет следующие особенности.

1. Разнородный сетевой трафик представляется как совокупность дискретных сообщений $S_{k,1}^{n_{Uk,1}}$, где $n_{Uk,1}$ – номер сообщения от последнего по порядку источника сообщений U_k к первому по порядку источнику сообщений U_1; k – количество узлов в информационной системе (ИС).

2. После приема сообщения с номером $n_{Uk-1,1}$ могут иметь место следующие события:

– с вероятностью p прием сообщения с номером $n_{Uk,1}$;

– с вероятностью q потеря сообщения с номером $n_{Uk,1}$.

Вероятность приёма следующего сообщения после приёма/передачи предыдущего сообщения, обозначим как $P_{k,1}^{n_{Uk,1}}$ – переходная вероятность приёма сообщения $S_{k,1}^{n_{U_k,1}}$ с порядковым номером $n_{Uk,1}$ после приёма сообщения $S_{k,1}^{n-1_{U_k,1}}$, отправленного от k-го узла к первому.

В соответствии с принятой моделью предельные (установившиеся или равновесные) вероятности состояний P_{sk} определяются из условия нормировки

$$\sum_{j=1}^{n_{Um,n}} P_{m,n}^{j} = 1,\ m = \overline{1,k},\ n = \overline{1,k}, \qquad (1)$$

согласно которому

$$P_{sk} = \frac{1-p/q}{1-(p/q)^{n-k+1}}\left(\frac{p}{q}\right)^{j-k} \quad j = k,\ldots,n. \quad (2)$$

Вероятности p и q связаны с заданным качеством сервиса в сети.

При выявлении аномального поведения и распознавании конкретных аномалий анализируются матрицы сигнатур и статистических показателей.

Зарегистрированные сигнатуры аномалий общим числом k обозначим $M = \{M_1, M_2, \ldots, M_k\}$. Сигнатурой M_1 описываются следующие характеристики и параметры:

– поле «адрес отправителя»;
– поле «адрес получателя»;
– поле «тип»;
– поле «данные»;
– поле «*CRC*»;
– непосредственно сами данные пакетов;
– время получения пакетов;
– время отправления пакетов;
– продолжительность сеанса связи в сети.

В состав сигнатуры M_2 могут входить:

– поле «адрес отправителя»;
– поле «адрес получателя»;
– непосредственно сами данные пакетов и т.д.

Совокупность n статистических показателей T сетевого трафика обозначим $T = \{T_1, T_2, \ldots, T_n\}$. К таким показателям относятся, например:

– среднее число входящих *IP*-пакетов в единицу времени;
– среднее число исходящих *IP*-пакетов в единицу времени;
– среднее число входящих *TCP*-пакетов в единицу времени;
– среднее число исходящих *TCP*-пакетов в единицу времени;
– среднее число входящих *UDP*-пакетов в единицу времени;
– среднее число исходящих *UDP*-пакетов в единицу времени;
– среднее время получения пакетов;
– среднее время отправления пакетов;
– средняя продолжительность сессии связи в сети;
– вероятности ошибок первого и второго рода.

Датчики системы обнаружения аномалий периодически передают информацию о состоянии защищаемого объекта. Если пороговый уровень β_{0I} превышен, принимается решение об обнаружении аномального поведения.

Датчики могут дать и ложную информацию, когда аномалия на самом деле не имеет места. Рассматриваемая задача относится к классу задач проверки статистических гипотез [2]. Для принятия решения, оптимального по некоторому критерию, необходимо выполнить соответствующие преобразования над принятой смесью полезного сигнала, мешающих сигналов и (помех) и шумов. Существует большое число статистических критериев оптимальности, однако известно [3], что все они приводят к стандартной процедуре вычисления отношения правдоподобия (ОП) или его монотонной функции (обычно вычисляют логарифм ОП) и сравнения результата с порогом β_{0i}:

$$L(x_i) = \frac{W(x_i / s_{1i}) >}{W(x_i / s_{0i}) <} \beta_{0i} \quad (3)$$

или

$$l(x_i) = \ln\frac{W(x_i / s_{1i}) >}{W(x_i / s_{0i}) <} \ln\beta_{0i}, \quad (3')$$

где $W(x_i / s_{1i})$ – условная плотность вероятности (ПВ) входной смеси при условии наличия полезного сигнала (попытка НСД); $W(x_i / s_{0i})$ – условная ПВ при условии отсутствия полезного сигнала (отсутствие попытки НСД, ложное срабатывание системы обнаружения вторжений).

По результатам обработки могут быть приняты следующие решения.

1. При использовании алгоритма с фиксированным объемом выборки $N_b = const$ и пороговым уровнем β_{ci}, который устанавливается в устройстве оптимальной обработки РЦН:

– $\beta_0 < \beta_{ci}$ – ложная тревога;

– $\beta_0 > \beta_{ci}$ – попытка НСД.

2. При использования алгоритма последовательного анализа с переменным объемом выборки $N_b = \text{var}$, нижним β_{ci} и верхним B_{ci} пороговыми уровнями, которые устанавливаются в устройстве оптимальной обработки:

– $\beta_0 < \beta_{ci}$ – ложная тревога;

– $\beta_0 > B_{ci}$ – попытка НСД;

– $\beta_{ci} < \beta_0 < B_{ci}$ – продолжение наблюдения.

Поскольку сигналы и помехи являются случайными величинами, объем выборки до принятия какого-либо решения также является случайной величиной и заранее не фиксируется. Интуитивно ясно, что чем меньше нижний порог и чем выше верхний порог, тем больше (в среднем) будет ожидаемый объем выборки и, соответственно, время наблюдения до момента принятия решения. Однако проблема выбора порогов является одной из центральных и наиболее сложных в теории проверки статистических гипотез. Известно [4], что выборе порогов в соответствии с неравенствами

$$\beta_{ci} \geq \min\left(\frac{1-P_{обн}}{1-P_{лт}}, \frac{P_{обн}}{P_{лт}}\right); \qquad (4)$$

$$B_{ci} \leq \max\left(\frac{1-P_{обн}}{1-P_{лт}}, \frac{P_{обн}}{P_{лт}}\right), \qquad (5)$$

где $P_{обн}$ – вероятность правильного обнаружения НСД; $P_{лт}$ – вероятность ложной тревоги, среднее время последовательного анализа является минимальным и меньшим, чем время анализа с выборками фиксированного объема (при тех же показателях качества обнаружения).

На практике интерес представляют условия высокой вероятности правильного обнаружения $P_{обн} > (0{,}9...0{,}99)$ и малой вероятности ложной тревоги $P_{лт} << 0{,}5$. Тогда неравенства (2 - 3) упрощаются:

$$\beta_{ci} \geq \left(\frac{1-P_{обн}}{1-P_{лт}}\right), \qquad (6)$$

$$B_{ci} \leq \left(\frac{1-P_{обн}}{1-P_{лт}}\right). \qquad (7)$$

На рис. 2 изображены графики зависимости отношения $\frac{B_{ci}}{\beta_{ci}}$ от вероятности правильного обнаружения $P_{обн}$ при постоянных вероятностях ложной тревоги $P_{лт}$.

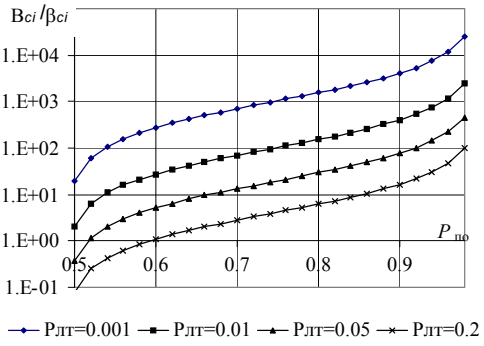

Рис. 2. Зависимости отношения порогов от вероятности правильного обнаружения

Выражения для математических ожиданий требуемых объемов выборок n до принятия решения имеют следующий вид [3]:

– по гипотезе H_0 (отсутствие аномалий)

$$m_0\{n|H_0\} = \frac{(1-P_{лт})\ln\frac{1-P_{обн}}{1-P_{лт}} + P_{лт}\ln\frac{P_{обн}}{P_{лт}}}{m_{10}}, \qquad (8)$$

– по гипотезе H_1 (наличие аномалий)

$$m_1\{n|H_1\} = \frac{(1-P_{обн})\ln\frac{1-P_{обн}}{1-P_{лт}} + P_{обн}\ln\frac{P_{обн}}{P_{лт}}}{m_{10}}, \qquad 9)$$

где m_{10} и m_{11} – математические ожидания корреляционных интегралов по гипотезам

H_0 и H_1 соответственно. Отношение $\frac{m_{11}}{m_{10}}$ можно трактовать как логарифм отношения (полезный сигнал)/(помеха + шум) по мощности.

На рис. 4 изображены графики зависимостей $m_0\{n \mid H_0\}$ и $m_1\{n \mid H_1\}$ от отношения $\frac{m_{11}}{m_{10}}$ при постоянных значениях $P_{обн}$ и $P_{лт}$.

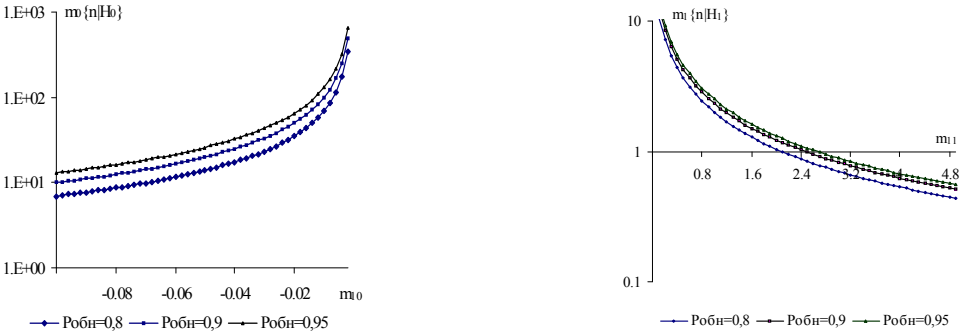

Рис. 3. Зависимости математических ожиданий объемов выборок от количественных соотношений обычных и аномальных состояний сети

Поскольку данные характеристики являются случайными и существенно меняются в процессе работы системы защиты, можно ожидать, что будут иметь место ситуации, когда процедура последовательного анализа окажется слишком длительной. Поэтому целесообразно использовать алгоритм усеченного последовательного анализа. При объеме выборки $n \le n_{\max}$, где $n_{\max}$ – заранее устанавливаемый максимально допустимый объем выборки, устанавливаются два порога, с которыми сравнивается ОП. Если же значение $n_{\max}$ достигнуто, а решение не принято, ОП сравнивается с одним порогом, как в алгоритме с фиксированным объемом выборки. Третий порог устанавливается внутри области между первым (нижним) и вторым (верхним) порогами. Метод оптимального выбора третьего порога пока не разработан. Поэтому рекомендуется подбирать его экспериментальным путем для конкретных условий работы системы защиты. Кроме того, необходимо предусмотреть возможность регулировки порогов в процессе работы.

Выводы

1. Судя по результатам расчетов, требуемый объем выборки слабо зависит от вероятности ложной тревоги, более существенно зависит от вероятности правильного обнаружения и сильнее всего зависит от статистических характеристик наблюдения обычных и аномальных состояний.

2. Система обнаружения аномалий, по существу, является автоматизированной человеко-машинной системой. Администратор сети анализирует состояния отдельных сегментов и сети в целом и, при необходимости, регулирует пороговые уровни процедур обнаружения.

3. В дальнейшем планируется исследовать асимптотические характеристики системы обнаружения аномалий с учетом их статистической взаимосвязи.

Список литературы

1. *F. Cohen.* Computer Viruses: theory and experiments // DOD/NBS 7th Conference on Computer Security (1984); Computers and Security, 1987. – vol. 6 №1. – P. 22–35.

2. *Леман Э.* Проверка статистических гипотез. – М.: Наука, 1979. – 408 с.

3. *Левин Б.Р.* Теоретические основы статистической радиотехники, книга вторая. – М.: Сов. радио, – 1968. – 504 с.

4. *Вальд А.* Последовательный анализ. – М.: Физматгиз, 1960. – 606 с.

Подано до редакції 10.03.10