Ariel Godinho
Felipe Vasconcelos

Modelo Epidemológico Modificado (SAIR) para vírus de computador

Brasil
2018, São Paulo

Ariel Godinho
Felipe Vasconcelos

Modelo Epidemológico Modificado (SAIR) para vírus de computador

Relatório Técnico para a resolução do Modelo Epidemológico Modificado (SAIR) para vírus de computador, como parte dos requisitos necessários para aprovação na disciplina Métodos Numéricos e Aplicações (MAP3122), dentro do Programa de Graduação em Engenharia de Computação.

Universidade de São Paulo – USP

Escola Politécnica

Programa de Graduação em Engenharia de Computação

MAP3122 – Métodos Numéricos e Aplicações

Brasil

2018, São Paulo

Lista de abreviaturas e siglas

SIR	Modelo epidemológico Suscetível-Infectado-Recuperado
SAIR	Modelo epidemológico Suscetível-Infectado-Recuperado-Vacinado

Lista de símbolos

S	População de computadores suscetíveis não-infectados
A	População de computadores vacinados não-infectados
I	População de computadores infectados
R	População de computadores removidos
N	Novos computadores adicionados a população definida inicialmente
μ	Coeficiente para a taxa de mortalidade não relacionada à infecção
β	Coeficiente de interação entre susceptíveis e infectados
α_{SA}	Coeficiente de interação entre susceptíveis e vacinados
α_{IA}	Coeficiente de interação entre infectados e vacinados
σ	Coeficiente para computadores consertados que voltam como susceptíveis
δ	Coeficiente para computadores removidos por inutilidade após infecção

Sumário

Parte I

Introdução

1 Introdução

1.1 Histórico

O primeiro trabalho acadêmico sobre a teoria dos programas de computador auto-replicantes foi feito em 1949 por John von Neumann, sendo posteriormente publicado como a "Teoria dos autômatos auto-reproduzidos". Em seu ensaio, von Neumann descreveu como um programa de computador poderia ser projetado para se reproduzir, sendo o design deste programa considerado o primeiro vírus de computador do mundo. Em 1972, Veith Risak, construindo diretamente sobre o trabalho de von Neumann sobre a auto-replicação, publicou seu artigo “Autômatos Auto Replicantes com Troca Mínima de Informações”. O artigo descreve um vírus totalmente funcional escrito em linguagem de programação para um sistema de computador. Em 1980, Jürgen Kraus escreveu sua tese de diploma “Auto-reprodução de programas”. Em seu trabalho, Kraus postulou que os programas de computador podem se comportar de forma semelhante à vírus biológicos.

1.2 Cenário Tecnológico

Infecções eram raras antes da internet se popularizar, e até os primeiros ataques pós-internet eram mais maliciosos do que criminosos. Com o aumento dos serviços financeiros on-line, a popularidade do comércio eletrônico e a presença de um mercado negro para informações de identificação pessoal, o malware tornou-se um grande negócio. Em retrospectiva, os primeiros dias de luta contra vírus eram quase pitorescos. As primeiras ferramentas eram basicamente verificadores de assinatura que procuravam mudanças em sistemas de arquivos ou aplicativos que correspondem a padrões conhecidos e, em seguida, sinalizavam ou bloqueavam a execução dos programas. Esta técnica ainda é usada hoje, mas tem algumas fraquezas fundamentais. Entre eles está a falha dos usuários em atualizar seus softwares regularmente e o fato de que leva tempo para catalogar todas as novas variantes de virus que são criadas todos os dias.
A arma mais comum de hoje é a verificação heurística de vírus, em que o código é analisado contra um conjunto de regras que indicam a presença de um vírus. Embora a abordagem heurística possa detectar a grande maioria dos vírus mais antigos, tem alguns dos mesmos pontos fracos que a abordagem da assinatura. Os desenvolvedores de vírus estão constantemente descobrindo novas maneiras de quebrar as regras, e é difícil para os fabricantes de software acompanharem.

1.3 Automodificação de Vírus

A maioria dos programas antivírus modernos tentam encontrar padrões de vírus dentro dos programas escaneando-os à procura de “assinaturas de vírus”. Infelizmente, o termo é enganador, na medida em que os vírus não possuem assinaturas únicas como os seres humanos tem. Essa "assinatura"de vírus é apenas uma seqüência de bytes que um programa antivírus busca porque é conhecido como parte do vírus. Um termo melhor seria "string de pesquisa". Diferentes programas antivírus irão empregar diferentes strings de pesquisa e, de fato, diferentes métodos de busca, na tentativa de identificar vírus. Se um scanner de vírus encontrar esse padrão em um arquivo, ele executará outras verificações para se certificar de que encontrou o vírus e não apenas uma sequência coincidente em um arquivo inocente, antes de notificar o usuário de que o arquivo está infectado. O usuário pode então excluir, ou (em alguns casos) "limpar"ou "curar"o arquivo infectado. Alguns vírus empregam técnicas que dificultam a detecção por assinatura, mas provavelmente não são impossíveis de serem detectados. Esses vírus modificam seu código em cada infecção. Ou seja, cada arquivo infectado contém uma variante diferente do vírus.

1.3.1 Vírus Cifrado

Um método de evadir a detecção de assinaturas é usar criptografia simples para cifrar (codificar) o corpo do vírus, deixando apenas o módulo de criptografia e uma chave criptográfica estática em texto claro que não muda de uma infecção para a próxima. Nesse caso, o vírus consiste em um pequeno módulo de decodificação e uma cópia criptografada do código do vírus. Se o vírus estiver criptografado com uma chave diferente para cada arquivo infectado, a única parte do vírus que permanece constante é o módulo de decodificação, que seria (por exemplo) anexado ao final. Neste caso, um scanner de vírus não pode detectar diretamente o vírus usando assinaturas, mas ainda pode detectar o módulo de decodificação, o que ainda possibilita a detecção indireta do vírus. Uma vez que estas seriam chaves simétricas armazenadas no hospedeiro infectado, é perfeitamente possível decifrar o vírus final, mas isso provavelmente não é necessário, uma vez que o código auto-modificador é de tamanha raridade que pode ser motivo para o antivírus pelo menos "marcar"o arquivo como suspeito.

Uma maneira antiga, mas compacta, seria o uso de operações aritmética como adição ou subtração e o uso de condições lógicas, como XOR, onde cada byte em um vírus é com uma constante, de modo que a operação XOR tenha apenas que ser repetida para decodificação. É suspeito que um código se modifique, então o código para criptografar/decriptografar pode ser parte da assinatura em muitas definições de vírus. Uma abordagem mais antiga e simples não usava uma chave, onde a criptografia consiste apenas em operações sem parâmetros, como incrementar e decrementar, rotação bit a bit, negação aritmética e operações NOT. Alguns vírus farão uso de um meio de criptografia dentro de um executável

em que o vírus é cifrado em determinados eventos, como o scanner de vírus que está sendo desabilitado para atualizações ou o computador sendo reiniciado. Isso é chamado de criptovirologia. Nos tempos indicados, o executável irá decifrar o vírus e executar suas rotinas ocultas, infectando o computador e às vezes desativando o software antivírus.

1.3.2 Código Polimórfico

O código polimórfico foi a primeira técnica que representou uma séria ameaça para os scanners de vírus. Assim como os vírus criptografados comuns, um vírus polimórfico infecta arquivos com uma cópia criptografada de si mesma, que é decodificada por um módulo de decriptografia. No caso de vírus polimórficos, no entanto, este módulo de decodificação também é modificado em cada infecção. Um vírus polimórfico bem escrito não tem partes que permanecem idênticas entre as infecções, tornando muito difícil a detecção diretamente usando "assinaturas". O software antivírus pode detectá-lo, decifrando o vírus usando um emulador, ou por análise de padrão estatístico do corpo do vírus criptografado. Para que o código seja polimórfico, o vírus deve ter um mecanismo polimórfico (também chamado de "motor mutante"ou "motor de mutação") em algum lugar em seu corpo criptografado.
Alguns vírus empregam código polimórfico de uma forma que se restringe significativamente a taxa de mutação do vírus. Por exemplo, um vírus pode ser programado para se modificar apenas um pouco ao longo do tempo, ou pode ser programado para abster-se de mutar quando infecta um arquivo em um computador que já contém cópias do vírus. A vantagem de usar esse código polimórfico lento é que torna mais difícil para profissionais e pesquisadores antivírus obter amostras representativas do vírus, porque os arquivos de "isca"que são infectados em uma execução normalmente contêm amostras idênticas ou similares do vírus. Isso tornará mais provável que a detecção pelo scanner de vírus não seja confiável e que algumas instâncias do vírus possam evitar a detecção.

1.3.3 Código Metamórfico

Para evitar ser detectado por emulação, alguns vírus se reescrevem completamente cada vez que estão para infectar novos executáveis. Os vírus que utilizam esta técnica são ditos ter código metamórfico. Para permitir o metamorfismo, é necessário um "motor metamórfico". Um vírus metamórfico geralmente é muito grande e complexo. Por exemplo, o W32/Simile consistiu em mais de 14.000 linhas de código de linguagem de montagem, 90% dos quais faz parte do mecanismo metamórfico.

1.4 Importância de Sistemas Seguros

O combate a programas maliciosos são de extrema importância na atualidade e continuarão a ser postos em consideração no futuro, pois o avanço de novas tecnologias e de programas maliciosos pedem que se tenha um esforço proporcional na busca por formas de proteger os sistemas atuais.
Assim dentro da necessidade de garantir a segurança e confiabilidade dos sistemas é fundamental entender como ocorre o processo de contaminação dos vírus em uma rede de computador. Pois assim, é possível ter um maior entendimento para criar soluções mais eficazes no combate a esses vírus.

1.5 Modelagem Inspirada na Área Biológica

O nome de vírus de computador não foi escolhido por acaso, mas sim por sua semelhança com o processo biológico de contaminação de doenças. O estudo de modelagem de vírus de computador como um vírus dentro da área biológica vem sendo pesquisada desde 1980, modelando essa dinâmica de modo análogo à contaminação de doenças reais. Essa modelagem pela área biológica pode ser feita por 2 cenários diferentes, uma micro – mais focada na proteção de um único sistema – e outro macro – mais focada no conjunto de sistemas conectados em uma rede única.
No cenário micro os vírus são vistos dentro do cenário com um único computador e de como se pode mitigar a contaminação desse computador por meios preventivos. Mas essa visão se torna limitada quando se pensa em uma rede de computadores, dificultando a visibilidade de possíveis soluções com uma atuação em massa para toda a rede de computadores. No cenário macro, é possível melhor entender essas forças que influenciam a dinâmica de propagação dos vírus em uma rede. Nessa visão, os estudos foram positivamente baseados em variações do modelo SIR, que modelam baseados em 3 estados para cada indivíduo (que nesse caso é um computador).
Os 3 estados são: Suscetível, Infectado e Recuperado. O estado suscetível caracteriza um sistema que pode ser infectado, mas que ainda não teve contato com o vírus. Já o estado de Infectado caracteriza os sistemas que estão infectados por esse vírus. Enquanto que o estado de Recuperado caracteriza os computadores que foram recuperados da infecção desse vírus.
O modelo usado para a solução do nosso problema será uma variante do modelo epidemiológico SIR, para representar indivíduos que possuam uma imunização contra o vírus em circulação na rede. O modelo possui o acrônimo de SAIR, em que A representar os indivíduos vacinados. Com esse modelo será possível aumentar a precisão da modelagem desse problema se comparado com a realidade dessa dinâmica. Na seção seguinte será descrito em maiores detalhes o funcionamento do Modelo SAIR proposto.

1.6 Sequência do Relatório

Nas sequência do relatório, na seção de Modelagem Matemática será exposto a modelagem do problema proposto usando o Modelo Epidemiológico Modificador SAIR.
Na seção de Metodologia Numérica iremos desenvolver como resolveremos o problema através das metodologias numéricas aprendidas na disciplina.
Na seção de Resultados será possível ver os resultados obtidos na solução do problema utilizando a metodologia explicada na seção de Metodologia Numérica.
Na seção de Conclusão será endereçado uma recapitulação de todo relatório evidenciando os pontos chave da resolução do problema proposto.

Parte II

Modelagem Matemática

2 Modelo SAIR

2.1 Descrição

O Modelo SIR, acrônimo para Modelo Suscetível-Infectado-Removido, baseia-se na interação entre 3 possíveis grupos para uma população de indivíduos – suscetíveis não-infectados, infectados e removidos por infecção ou não – frente a uma doença em questão. No caso deste trabalho, a doença é um vírus de computador, e consequentemente, os indivíduos são os computadores. Para uma aproximação mais realista do comportamento de um vírus entre os computadores, esse modelo modificado propõe mais um novo estado para essa população: o estado de indivíduos vacinados não-infectados (traduzido de "antidotal"do inglês). No nosso caso, a vacina significaria que o computador possui um antivírus capaz de identificar o vírus e assim, não pode ser infectado. A seguir, é possível ver a descrição para o Modelo SAIR proposto, assim como a explicação de cada variável. Na Figura 1 podemos ver a interação entre as quatro populações e seus coeficientes de interação e na equação 2.1 podemos ver o modelo completo.

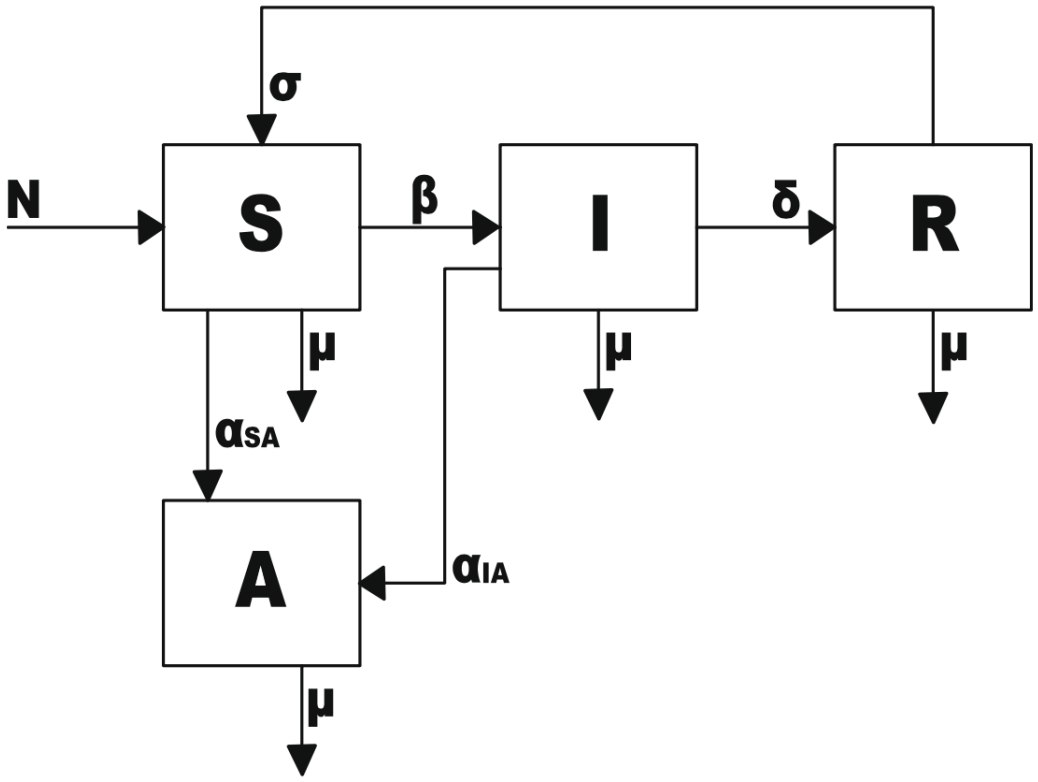

Figura 1 – Interação entre os indivíduos no modelo SAIR.

S População de computadores suscetíveis não-infectados

A População de computadores vacinados não-infectados

I População de computadores infectados

R População de computadores removidos

N Novos computadores adicionados a população definida inicialmente

μ Coeficiente para a taxa de mortalidade não relacionada à infecção

β Coeficiente de interação entre susceptíveis e infectados

α_{SA} Coeficiente de interação entre susceptíveis e vacinados

α_{IA} Coeficiente de interação entre infectados e vacinados

σ Coeficiente para computadores consertados que voltam como susceptíveis

δ Coeficiente para computadores removidos por inutilidade após infecção

$$\begin{cases} \dot{S} = N - \alpha_{SA}SA - \beta SI - \mu S + \sigma R \\ \dot{I} = \beta SI - \alpha_{IA}AI - \delta I - \mu I \\ \dot{R} = \delta I - \sigma R - \mu R \\ \dot{A} = \alpha_{SA}SA + \alpha_{IA}AI - \mu A \end{cases} \tag{2.1}$$

2.2 Equações e Modelagem

Supondo que a velocidade de propagação do vírus é mais rápida que a entrada de novos computadores na rede e da obsolência de computadores existentes na rede, usaremos $N = 0$ e $\mu = 0$. A partir dessa suposição e analisando o diagrama da Figura 1, vemos que o sistema que era aberto se torna um sistema fechado, ou seja, a população ao longo do tempo se mantém constante. O que simplifica o modelo para o apresentado na equação 2.2:

Nota-se que T = S + I + R + A, logo podemos eliminar uma das equações, tomar T = 100 e usar S, I, R e A como porcentagens.

$$\begin{cases} \dot{S} = -\alpha_{SA}SA - \beta SI + \sigma R \\ \dot{I} = \beta SI - \alpha_{IA}AI - \delta I \\ \dot{R} = \delta I - \sigma R \\ \dot{A} = \alpha_{SA}SA + \alpha_{IA}AI \end{cases} \tag{2.2}$$

2.2.1 Pontos de Equilíbrio

Um ponto de equilíbrio ocorre quando a proporção entre os indivíduos não muda com a variação do tempo, significando que o sistema convergiu para um ponto no sistema de coordenadas S x I x R. Os pontos de equilíbrio não-endêmicos são $P_1 = (S, I, R, A) = (0, 0, 0, T)$ e $P_2 = (S, I, R, A) = (T, 0, 0, 0)$ que indica um equilíbrio sem indivíduos infectados. Um Ponto de Equilíbrio Endêmico ocorre quando a proporção entre os indivíduos não muda e temos $I \neq 0$. Neste ponto, denominado P3, temos que:

$$\begin{cases} S = \dfrac{\delta}{\beta} \\ I = \dfrac{T - \dfrac{\delta}{\beta}}{1 + \dfrac{\delta}{\sigma}} \\ R = \dfrac{T - \dfrac{\delta}{\beta}}{1 + \dfrac{\sigma}{\delta}} \\ A = 0 \end{cases} \tag{2.3}$$

Para tal ponto existir, é necessário que $T < \dfrac{\delta}{\beta}$, portanto os valores iniciais escolhidos para a equação serão:

$$\begin{aligned} T &= 100 \\ S_0 &= 95 \\ A_0 &= 0 \\ I_0 &= 5 \\ R_0 &= 0 \\ N &= 0 \\ \mu &= 0 \\ \beta &= 0.1 \end{aligned}$$

$$\alpha_{SA} = 0.025$$

$$\alpha_{IA} = 0.25$$

$$\delta = 0.8$$

$$\gamma = 9$$

Assim podemos calcular o ponto P3 usando a equação 2.3:

$$\begin{cases} S = 90 \\ I = 0.82 \\ R = 9.18 \\ A = 0 \end{cases} \tag{2.4}$$

Logo, $P_3 = (90, 0.82, 9.18, 0)$.

Caso A seja diferente de 0, a equação convergirá para $P_1(0, 0, 0, T)$