uc3m | Universidad Carlos III de Madrid

Master Degree in Cybersecurity

2020-2021

Master Thesis

"SIMULATING VIRUS PROPAGATION IN BROAD COMPUTER NETWORKS"
Supplementary Documentation

Christian Durán González

David Expósito Singh

Madrid, 30th of September 2021

1. INTRODUCTION

NetSim is a scalable, parallel agent-based simulator based on MPI that has been implemented in order to represent the transmission of an email worm between companies of different sizes considering email user behaviors, worm characteristics, antivirus installation and network conditions.

The objective of this document, which complements the information in Master Thesis "SIMULATING VIRUS PROPAGATION IN BROAD COMPUTER NETWORKS", is to explain data management. Given that NetSim implements several modules that reproduce the different aspects which have an impact on the virus propagation. For this, different heterogeneous sources and data types are taken as input by mapping them to the different parameters of the agent model. In addition, we will briefly introduce the state diagram for the nodes that have an antivirus installed.

Figure 1.1 shows the different stages involved in NetSim simulation. Its first two stages are described in section 2. During the first stage, heterogeneous input data is obtained from different databases and papers. In the second stage, these data are processed using multiple technologies.

The third stage corresponds to the NetSim simulation, which includes models that reproduce the network where the worm is transmitted, taking into account human behavior, the characteristics of the worm, the periods of the states and the transition probabilities.

The network topology model reproduces a heterogeneous contact network where the computing devices as nodes and the connections between the user and the nodes in the user's contact list as edges are represented. The model takes into account the social habits of users, since it allows work, family and leisure connections.

In addition, Figure 1.1 includes a fourth stage for post-processing the results and a fifth for generating statistics and graphs that summarize the simulation output.

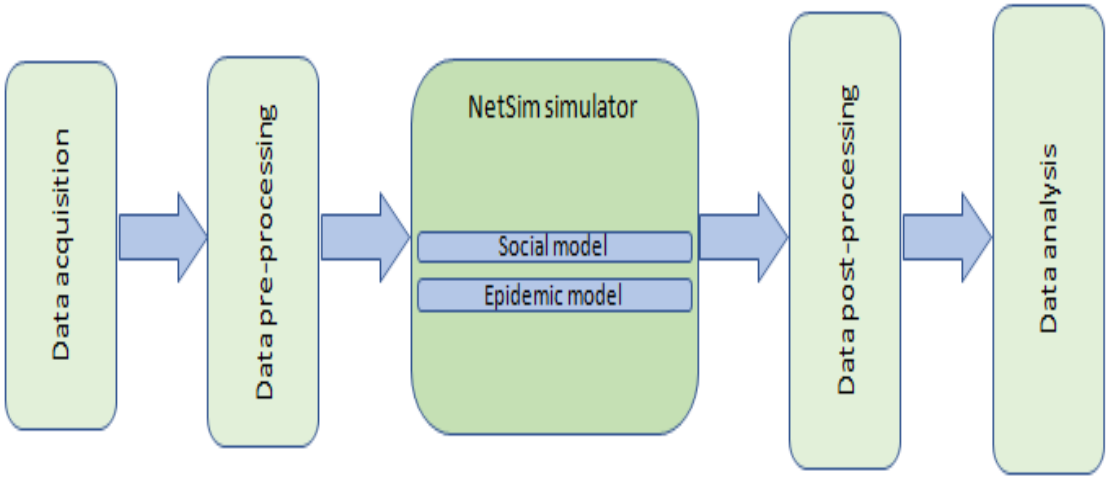

Fig. 1.1. Stages involved in NetSim simulations.

2. DATA FLOW

This section describes the data used, as well as how it was generated and processed. The Figure2.1 shows the different data sources involved in a simulation and how they interact with the different software components. NetSim consists of two main software pieces combined with many auxiliary programs. The first component is the generation of the scenario composed of nodes that are grouped into companies. Within the data sources we find the information obtained from Enron Email Corpus and contact matrices, extracted from public surveys, are used to provide statistical information of the average number of contacts between individuals of certain age ranges. Another data source is the INE [1] information which provides the demographic data used by the simulator to find out the statistics on Spanish citizens and their social and work behavior. The NetSim generator is an MPI program written in C that creates the scenario through the characteristics of the users (age, profession, etc.). It is called social fabric, and is used as input in the scenario simulation (lower part of Figure 2.1). It is stored as sparse arrays, in which each node is a node and each edge is a time-dependent interaction with other node. Furthermore, we consider the epidemiological model of email worm whose parameters related to human behavior were taken from the existing literature. The previous data is used by NetSim to perform the scenario simulation. The simulator output is a collection of trace files with the state of each node for each simulated time step.

2.1. Scenario generation

Demographic data is a set of data that defines the characteristics of users. To obtain the data for Spain, we use the information from the Spanish National Statistics Institute (INE) that indicates the statistics of the main labor sectors, where each one has a different behavior: industry, building, catering, services, security forces, education, front-line health, non front-line health, social-health and transport. In addition, the data provided by the INE on the household size is used for the modeling of family contacts, which indicates the statistics of the members who are in the households. Demographic data also includes the minimum, average and maximum number of workers that a company can have in Spain. Population-mixing data is used to generate the social pattern where realism is decisive:

- Social network graphs obtained from the Enron Email Corpus graph [2] (70,578 nodes and 312,620 edges) are used for generating the work and informal meetup contacts. A variation [3] of the Random Walk algorithm [4] is implemented in order to generate scaled sub-graphs with an specific average connectivity <k>provided as input argument, whose value is usually obtained from contact arrays.

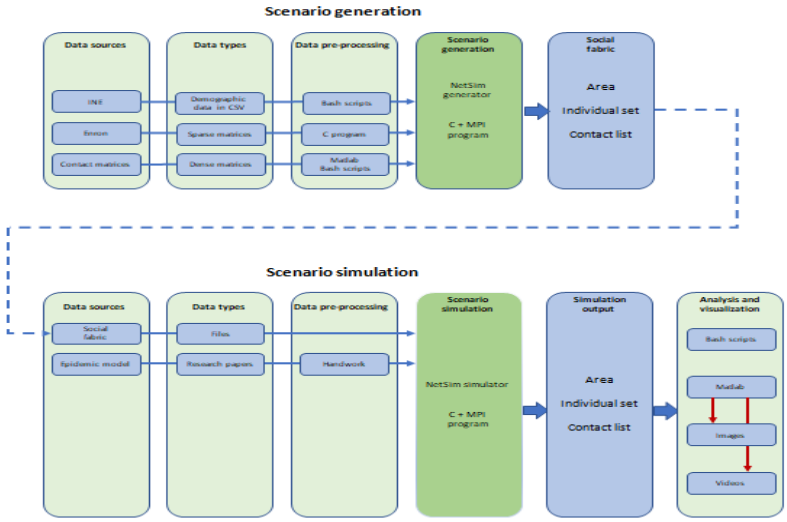

Fig. 2.1. NetSim dataflow.

- Contact matrices [5] are dense matrices in which each element Ai,j represents average number of daily interactions between individuals of ages i and j. The contact matrix repository includes data for various countries and some regions of these countries, including sub-contact matrices for school, work, and community contacts. These contact matrices were processed using Matlab and Bash scripts. The work contact matrix was used to define the age distribution of the work and the leisure contacts were created through the community contact.

2.2. Scenario simulation

The epidemic model, originated after the information obtained from the literature published in recent years, takes into account the following characteristics:

- We assume a verification time of the user's email (called T_i) as a stochastic process, and the probability of opening the attached file (referenced as C_i) will follow a normal distribution. Users are assumed to have independent behaviours which allows us to take on that the email verification periods and the probability of opening it behave as independent Normal distributions.
- We evaluate the use of antivirus by modelling the percentage of nodes which have an antivirus installed.
- Those nodes that have an antivirus installed will have greater protection when opening an attachment.
- The worm can be patched by itself and thus reduce the probability of reinfecting when opening an attachment containing this worm. This patching can be partial or total, in which case the probability of being reinfected would be zero.

3. ANTIVIRUS STATES

The master thesis presents the states through which a node without installed antivirus transits. However, nodes with antivirus have the same states:

- Latent stage: the node is in this state every time the user checks if an email worm has been received. If the user opens the email, the model of the node will transition to the pre-infection state and if not, it will return to susceptible.
- Pre-infection stage: the node is infected, but we model the time it takes to propagates to the user's contacts. After this state, the node changes to the broadcast infection state.
- Broadcast infection stage: at the end of this stage, the node may be immune, isolated, died or return to susceptible
- Isolation stage: the node is isolated from the network for a certain time because an infection or abnormal is detected. At the end of this stage the node model will be able to transition to the immunization, death or susceptible state.
- Immunization stage: the node will not be able to be infected again, once the broadcast infection or isolated state have been completed.
- Death stage: a node dies due to the infection or other reasons which halts the computer equipment, subsequently transitioning to a susceptible state.

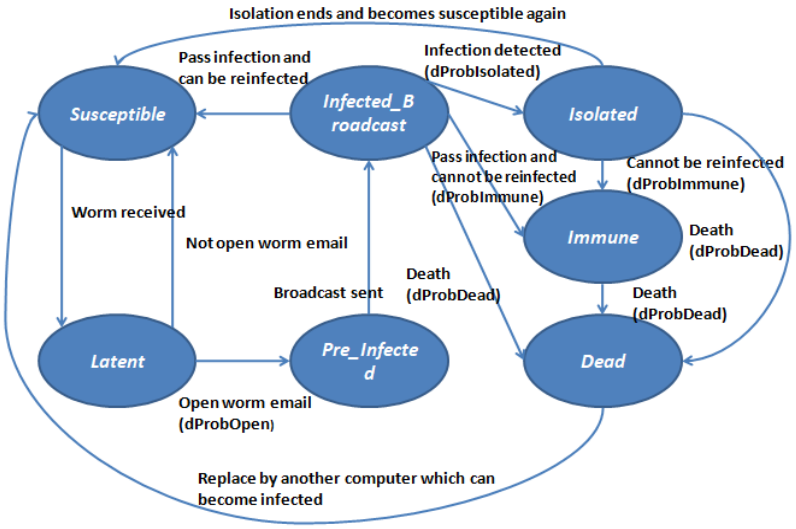

Fig. 3.1. Compartmental model. It consists of the following states: SUSCEPTIBLE, LATENT, PRE-INFECTED (pre-infection stage), INFECTED BROADCAST (broadcast infection stage), ISOLATED (isolation stage), IMMUNE (immunization stage) and DEAD (death stage). The edges show the transition normalized probabilities between the stages. Note that between the pre-infection state and the infected broadcast state there is no indicated probability because that transition always takes place. In addition, we create the 'treated' states between which the nodes with antivirus installed pass because their periods are different from those of the nodes that do not have antivirus.

4. CONCLUSION

This document describes the data management of NetSim, an agent-based simulator for email worm propagation through the use of different, complex and heterogeneous data sources. MPI is used by NetSim in order to execute the C simulator in parallel on multiple compute nodes. We use Python, Bash scripts and Matlab for the information processing and the elaboration of graphs and statistics. In this way, several processing scripts can be simultaneously executed in order to speed-up the pre-processing and post-processing stages.

BIBLIOGRAFÍA

[1] *National Statistics Institute (INE)*, Ministry of Economic Affairs y Digital Transformation (MINECO), http://www.ine.es, 2021.

[2] W. Cukierski, *The Enron Email Dataset*, https://www.kaggle.com/wcukierski/enron-email-dataset, 2015.

[3] M. Guzmán-Merino et al., "Assessing population-sampling strategies for reducing the COVID-19 incidence," *Preprint*, 2021.

[4] L. L., "Random Walks on Graphs: A Survey," *BOLYAI SOCIETY MATHEMATICAL STUDIES*, vol. 2, p. 46, 1993. [En línea]. Disponible en: https://web.cs.elte.hu/~lovasz/erdos.pdf.

[5] M. Chinazzi y D. Mistry, *Mixing Patterns*, https://github.com/mobs-lab/mixing-patterns, 2021.