Epigrass: a tool to study disease spread in complex networks.

Flávio Codeço Coelho*, Oswaldo Gonçalves Cruz , Cláudia Torres Codeço

Programa de Computação Científica - Fundação Oswaldo Cruz - Av. Brasil,4365 - Rio de Janeiro - RJ - Brasil - 21045-900

Email: Flávio Codeço Coelho*- fccoelho@fiocruz.br; Oswaldo Gonçalves Cruz - oswaldo@fiocruz.br; Cláudia Torres Codeço - codeco@fiocruz.br;

*Corresponding author

Abstract

Background: The construction of complex statial simulation models such as those used in network epidemiology, is a daunting task due to the large amount of data involved in their parameterization. Such data, which frequently resides on large geo-referenced databases, has to be processed and assigned to the various components of the model. All this just to construct the model, then it still has to be simulated and analyzed under different epidemiological scenarios. This workflow can only be achieved efficiently by computational tools that can automate most if not all these time-consuming tasks. In this paper, we present a simulation software, Epigrass, aimed to help designing and simulating network-epidemic models with any kind of node behavior.

Results: A Network epidemiological model representing the spread of a directly transmitted disease through a bus-transportation network connecting mid-size cities in Brazil. Results show that the topological context of the starting point of the epidemic is of great importance from both control and preventive perspectives.

Conclusions: Epigrass is shown to facilitate greatly the construction, simulation and analysis of complex network models. The output of model results in standard GIS file formats facilitate the post-processing and analysis of results by means of sophisticated GIS software.

Background

Epidemic models describe the spread of infectious diseases in populations. More and more, these models are being used for predicting,understanding and developing control strategies. To be used in specific contexts, modeling approach have shifted from "strategic models" (where a caricature of real processes is modeled in order to emphasize first principles)to "tactical models" (detailed representations of real situations). Tactical models are useful for cost-benefit and scenario analyzes. Good examples are the foot-and-mouth epidemic models for UK, triggered by the need of a response to the 2001 epidemic [1,2] and the simulation of pandemic flu in different scenarios helping authorities to choose among alternative intervention strategies [3, 4].

In realistic epidemic models, a key issue to consider is the representation of the contact process through which a disease is spread, and networks have arisen as good candidate [5]. This has led to the de-

velopment of "network epidemic models". Network is a flexible concept that can be used to describe, for example, a collection of individuals linked by sexual partnerships [6], a collection of families linked by sharing workplaces/ schools [7], a collection of cities linked by air routes [8]. Any of these scales may be relevant to the study and control of disease spread [9].

Networks are made of nodes and their connections. One may classify network epidemic models according to node behavior. One example would be a classification based on the states assumed by the nodes: networks with discrete-state nodes have nodes characterized by a discrete variable representing its epidemiological status (for example, susceptible, infected, recovered). The state of a node changes in response to the state of neighbor nodes, as defined by the network topology and a set of transmission rules. Networks with continuous-state nodes, on the other hand, have nodes with state described by dynamic variables whose value is a function of the history of the node and its neighbors. The importance of the concept of neighborhood on any kind of network epidemic model stems from its large overlap with the concept of transmission. In network epidemic models, transmission either defines or is defined/constrained by the neighborhood structure. In the latter case, a neighborhood structure is given a priori which will influence transmissibility between nodes.

The construction of complex simulation models such as those used in network epidemic models, is a daunting task due to the large amount of data involved in their parameterization. Such data frequently resides on large geo-referenced databases. This data has to be processed and assigned to the various components of the model. All this just to construct the model, then it still has to be simulated, analyzed under different epidemiological scenarios. This workflow can only be achieved efficiently by computational tools that can automate most if not all of these time-consuming tasks.

In this paper, we present a simulation software, Epigrass, aimed to help designing and simulating network-epidemic models with any kind of node behavior. Without such a tool, implementing network epidemic models is not a simple task, requiring a reasonably good knowledge of programming. We expect that this software will stimulate the use and development of networks models for epidemiological purposes.

The paper is organized as following: first we describe the software and how it is organized with a brief overview of its functionality. Then we demonstrate its use with an example. The example simulates the spread of a directly transmitted infectious disease in Brazil through its transportation network. The velocity of spread of new diseases in a network of susceptible populations depends on their spatial distribution, size, susceptibility and patterns of contact. In a spatial scale, climate and environment may also impact the dynamics of geographical spread as it introduces temporal and spatial heterogeneity. Understanding and predicting the direction and velocity of an invasion wave is key for emergency preparedness.

Epigrass

Epigrass is a platform for network epidemiological simulation and analysis. It enables researchers to perform comprehensive spatio-temporal simulations incorporating epidemiological data and models for disease transmission and control in order to create complex scenario analyses. Epigrass is designed towards facilitating the construction and simulation of large scale metapopulational models. Each component population of such a metapopulational model is assumed to be connected through a contact network which determines migration flows between populations. This connectivity model can be easily adapted to represent any type of adjacency structure.

Epigrass is entirely written in the Python language, which contributes greatly to the flexibility of the whole system due to the dynamical nature of the language. The geo-referenced networks over which epidemiological processes take place can be very straightforwardly represented in a object-oriented framework. Consequently, the nodes and edges of the geographical networks are objects with their own attributes and methods (figure 1).

Once the archetypal node and edge objects are defined with appropriate attributes and methods, then a code representation of the real system can be constructed, where nodes (representing people or localities) and contact routes are instances of node and edge objects, respectively. The whole network is also an object with its own set of attributes and methods. In fact, Epigrass also allows for multiple edge sets in order to represent multiple contact networks in a single model.

These features leads to a compact and hierarchi-

cal computational model consisting of a network object containing a variable number of node and edge objects. It also does not pose limitations to encapsulation, potentially allowing for networks within networks, if desirable. This representation can also be easily distributed over a computational grid or cluster, if the dependency structure of the whole model does not prevent it.

For the end-user, this representation is not an obstacle since it reflects the natural structure of the real system. Even after the model is converted into a code object, all of its component objects remain accessible to one another, facilitating the exchange of information between all levels of the model, a feature the user can easily include in his/her custom models.

Nodes and edges are dynamical objects in the sense that they can be modified at runtime altering their behavior in response to user defined events. In Epigrass it is very easy to simulate any dynamical system embedded in a network. However, it was designed with epidemiological models in mind. This goal led to the inclusion of a collection of built-in epidemic models which can be readily used for the intra-node dynamics. Epigrass users are not limited to basing their simulations on the built-in models. User-defined models can be developed in just a few lines of Python code. All simulations in Epigrass are done in discrete-time. However, custom models may implement finer dynamics within each time step, by implementing ODE models at the nodes, for instance.

The Epigrass system is driven by a graphical user interface(GUI), which handles several input files required for model definition and manages the simulation and output generation.

At the core of the system lies the simulator. It parses the model specification files, contained in a text file (.epg file), and builds the network from site and edge description files (comma separated values text files, CSV). The simulator then builds a code representation of the entire model, simulates it,and stores the results in the database (figure 2) or in a couple of CSV files. This output will contain the full time series of the variables in the model. Additionally, a map layer (in shapefile and KML format) is also generated with summary statitics for the model.

The results of an Epigrass simulation can be visualized in different ways. A 3D map with an animation of the resulting timeseries is available directly through the GUI (figure 9). Other types of static visualizations can be generated through GIS software from the shapefiles generated. The KML file can also be viewed in Google Earth™ or Google Maps™ (figure 3).

Epigrass also includes a report generator module which is controlled through a parameter in the ".epg" file. Epigrass is capable of generating PDF reports with summary statistics from the simulation. This module requires a LaTeXinstallation to work. Reports are most useful for general verification of expected model behavior and network structure. However the LaTeXsource files generated by the module may serve as templates that the user can edit to generated a more complete document.

Building a model in Epigrass is very simple, especially if the user chooses to use one of the built-in models. Epigrass includes 20 diferent epidemic models ready to be used (See manual for built-in models description).

To run a network epidemic model in Epigrass, the user is required to provide three separate text files (Optionally, also a shapefile with the map layer):

1. Node-specification file: This file can be edited on a spreadsheet and saved as a csv file. Each row is a node and the columns are variables describing the node.

2. Edge-specification file: This is also a spreadsheet-like file with an edge per row. Columns contain flow variables.

3. Model-specification file: Also referred to as the ".epg" file. This file specifies the epidemiological model to be run at the nodes, its parameters, flow model for the edges, and general parameters of the simulation.

The ".epg" file is normally modified from templates included with Epigrass. Nodes and edges files on the other hand, have to be built from scratch for every new network. Details of how to construct these files, as well as examples, can be found in the documentation accompanying the software, which is available at http://epigrass.sourceforge.net.

Methods

Example Model

In the example application, the spread of a respiratory disease through a network of cities connected by bus transportation routes is analyzed.

The epidemiological scenario is one of the invasion of a new influenza-like virus. One may want to simulate the spread of this disease through the country by the transportation network to evaluate alternative intervention strategies (e.g. different vaccination strategies). In this problem, a network can be defined as a set of nodes and links where nodes represent cities and links represents transportation routes. Some examples of this kind of model are available in the literature [8, 10].

One possible objective of this model is to understand how the spread of such a disease may be affected by the point-of-entry of the disease in the network. To that end, we may look at variables such as the speed of the epidemic, number of cases after a fixed amount of time, the distribution of cases in time and the path taken by the spread.

The example network was built from 76 of largest cities of Brazil (>= 100k habs). The bus routes between those cities formed the connections between the nodes of the networks. The number of edges in the network, derived from the bus routes, is 850. These bus routes are registered with the National Agency of Terrestrial Transportation (ANTT) which provided the data used to parameterize the edges of the network.

The model

The epidemiological model used consisted of a metapopulation system with a discrete-time SEIR model (Eq. 1). For each city, S_t is the number of susceptibles in the city at time t, E_t is the number of infected but not yet infectious individuals, I_t is the number of infectious individuals resident in the locality, N is the population residing in the locality (assumed constant throughout the simulation), and n_t is the number of individuals visiting the locality , Θ_t is the number of visitors who are infectious. The parameters used were taken from Lipsitch et al. (2003) [11] to represent a disease like SARS with an estimated basic reproduction number (R_0) of 2.2 to 3.6 (Table 1).

$$\begin{aligned} S_{t+1} &= S_t - \beta S_t \frac{(I_t + \Theta)^\alpha}{N_t + n_t} \\ E_{t+1} &= (1-e)E_t + \beta S_t \frac{(I_t + \Theta_t)^\alpha}{N_t + n_t} \\ I_{t+1} &= eE_t + (1-r)I_t \\ R_{t+1} &= N_t - (S_{t+1} + I_{t+1} + E_{t+1}) \end{aligned} \quad (1)$$

To simulate the spread of infection between cities, we used the concept of a "forest fire" model [12]. An infected individual, traveling to another city, acts as a spark that may trigger an epidemic in the new locality. This approach is based on the assumption that individuals commute between localities and contribute temporarily to the number of infected in the new locality, but not to its demography. Implications of this approach are discussed in Grenfell et al (2001) [12].

The number of individuals arriving in a city (n_t) is based on annual total number of passengers arriving trough all bus routes leading to that city as provided by the ANTT (www.antt.gov.br). The annual number of passengers is used to derive an average daily number of passengers simply by dividing it by 365.

Stochasticity is introduced in the model at two points: the number of new cases is draw from a Poisson distribution with intensity $= \frac{(I_t+\Theta_t)^\alpha}{N_t+n_t}$ and the number of infected individuals visiting i is modelled as binomial process:

$$\begin{aligned} \Theta_t &= \sum_k \theta_{k,t} \text{ for all k neighbors} \\ \theta_{k,t} &\sim Binomial\left(n, \frac{I_{k,t-\delta}}{N_k}\right) \end{aligned}$$

where n is the total number of passengers arriving from a given neighboring city, $I_{k,t}$ and N_k are the current number of infectious and population size of city k, respectively. δ is the delay associated with the duration of each bus trip. The delay δ was calculated as the number of days (rounded down) that a bus, traveling at an average speed of $60km/h$, would take to complete a given trip. The lengths in kilometers of all bus routes were also obtained from the ANTT.

The files with this model's definition(the sites, edges and ".epg" files) are available as part of the supplementary material for this article.

Analysis

In the context of this model, we say that a city has become infected when it displays its first authoctomous case. The exposure of a city to invasion is defined as inversely related to the time elapsed from the beginning of the epidemic until the arrival of the first case in the city.

To determine the importance of the point of entry in the outcome of the epidemic, the model was run 500 times, randomizing the point of entry of the virus. The seeding site was chosen with a probability proportional to the log_{10} of their population size. These replicates were run using Epigrass' built-in support for repeated runs with the option of randomizing seeding site. For every simulation, statistics about each site such as centrality, betweeness, and time it got infected were saved.

The time required for the epidemic to infect 50% of the cities was chosen as a global index to network susceptibility to invasion.

Except for population size, all other epidemiological parameters were the same for all cities, that is, disease transmissibility and recovery rate. Some positional features of each node were derived from the graph: *Centrality*, which is the inverse of the sum of the shortest paths from a node to all other nodes in the graph; *Betweeness*, which is the number of times a node figures in the the shortest path between any other pair of nodes; and *Degree*, which is the number of edges connected to a node.

In order to analyze the path of the epidemic spread we recorded which cities provided the infectious cases which were responsible for the infection of each city. if more than on source of infection exists, Epigrass selects the city which contributed with the largest number of infectious individuals at that time-step, as the most likely infector. At the end of the simulation Epigrass generates a file with the dispersion tree in graphML format, which can be read by a variety of graph plotting programs the generate the graphic seen on figure 5.

Results and Discussion

Effects of the point of entry:

The spread speed of the epidemic, measured as the time taken to infect 50% of the cities, was found to be influenced by the centrality and degree of the entry node (figure 4).

The dispersion tree corresponding to the epidemic, is greatly influenced by the degree of the point of entry of the disease in the network. Figure 5 shows the tree for the dispersion from the city of Salvador.

Vaccination strategies must take into consideration network topology. Figures 6 and 7 show cost benefit plots for three vaccination strategies investigated: *Uniform vaccination*, *top-3 degree sites only* and *top-10 degree sites only*. Vaccination of higher order sites offer cost/benefit advantages only in scenarios where the disease enter the network through one of these sites.

Conclusions

Epigrass facilitates greatly the construction, simulation and analysis of complex network models. The output of model results in standard GIS file formats facilitates the post-processing and analysis of results by means of sophisticated GIS software.

Besides invasion, network epidemiological models can also be used to understand patterns of geographical spread of endemic diseases [13–16]. Many infectious diseases can only be maintained in a endemic state in cities with population size above a threshold, or under appropriate environmental conditions(climate, availability of a reservoir, vectors, etc). The variables and the magnitudes associated with endemicity threshold depends on the natural history of the disease [17]. Theses magnitudes may vary from place to place as it depends on the contact structure of the individuals.Predicting which cities are sources for the endemicity and understanding the path of recurrent traveling waves may help us to design optimal surveillance and control strategies.

Authors contributions

Flavio Codeço Coelho contributed with the software development, model definition and analysis as well as general manuscript conception and writing. Cláudia Torres Codeço contributed with model definition and implementation, as well as with writing the manuscript. Oswaldo Cruz, contributed with data analysis and writing the manuscript.

Acknowledgements

The Authors would like to thank the Brazilian Research Council (CNPq) for financial support to the

authors.

References

1. Keeling MJ, Woolhouse MEJ, May RM, Davies G, Grenfell BT: **Modelling vaccination strategies against foot-and-mouth disease.** *Nature* 2003, **421**(6919):136–142, [http://dx.doi.org/10.1038/nature01343].

2. Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, Woolhouse MEJ, Grenfell BT, Keeling MJ: **Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK.** *Nature* 2006, **440**(7080):83–86, [http://dx.doi.org/10.1038/nature04324].

3. Longini IM, Halloran ME: **Strategy for distribution of influenza vaccine to high-risk groups and children.** *Am J Epidemiol* 2005, **161**(4):303–306, [http://dx.doi.org/10.1093/aje/kwi053].

4. Longini IM, Halloran ME, Nizam A, Yang Y: **Containing pandemic influenza with antiviral agents.** *Am J Epidemiol* 2004, **159**(7):623–633.

5. Parham PE, Ferguson NM: **Space and contact networks: capturing the locality of disease transmission.** *J R Soc Interface* 2006, **3**(9):483–493, [http://dx.doi.org/10.1098/rsif.2005.0105].

6. Handcock MS, Jones JH: **Interval estimates for epidemic thresholds in two-sex network models.** *Theor Popul Biol* 2006, **70**(2):125–134, [http://dx.doi.org/10.1016/j.tpb.2006.02.004].

7. Meyers LA, Newman MEJ, Martin M, Schrag S: **Applying network theory to epidemics: control measures for Mycoplasma pneumoniae outbreaks.** *Emerg Infect Dis* 2003, **9**(2):204–210.

8. Grais RF, Ellis JH, Glass GE: **Assessing the impact of airline travel on the geographic spread of pandemic influenza.** *Eur J Epidemiol* 2003, **18**(11):1065–1072.

9. Pourbohloul B, Meyers LA, Skowronski DM, Krajden M, Patrick DM, Brunham RC: **Modeling control strategies of respiratory pathogens.** *Emerg Infect Dis* 2005, **11**(8):1249–1256.

10. Longini IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DAT, Halloran ME: **Containing pandemic influenza at the source.** *Science* 2005, **309**(5737):1083–1087, [http://dx.doi.org/10.1126/science.1115717].

11. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M: **Transmission dynamics and control of severe acute respiratory syndrome.** *Science* 2003, **300**(5627):1966–1970, [http://dx.doi.org/10.1126/science.1086616].

12. Grenfell BT, Bjørnstad ON, Kappey J: **Travelling waves and spatial hierarchies in measles epidemics.** *Nature* 2001, **414**(6865):716–723, [http://dx.doi.org/10.1038/414716a].

13. Cummings DAT, Irizarry RA, Huang NE, Endy TP, Nisalak A, Ungchusak K, Burke DS: **Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand.** *Nature* 2004, **427**(6972):344–347, [http://dx.doi.org/10.1038/nature02225].

14. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N: **Modelling disease outbreaks in realistic urban social networks.** *Nature* 2004, **429**(6988):180–184, [http://dx.doi.org/10.1038/nature02541].

15. Raffy M, Tran A: **On the dynamics of flying insects populations controlled by large scale information.** *Theor Popul Biol* 2005, **68**(2):91–104, [http://dx.doi.org/10.1016/j.tpb.2005.03.005].

16. Riley S: **Large-scale spatial-transmission models of infectious disease.** *Science* 2007, **316**(5829):1298–1301, [http://dx.doi.org/10.1126/science.1134695].

17. Keeling MJ, Grenfell BT: **Disease extinction and community size: modeling the persistence of measles.** *Science* 1997, **275**(5296):65–67.

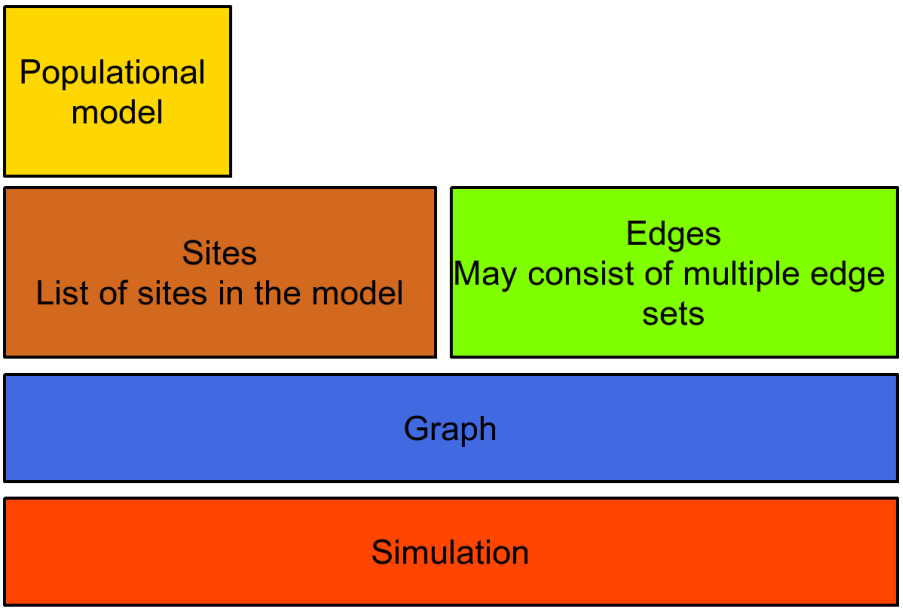


Figure 1: Architecture of the Epigrass system.

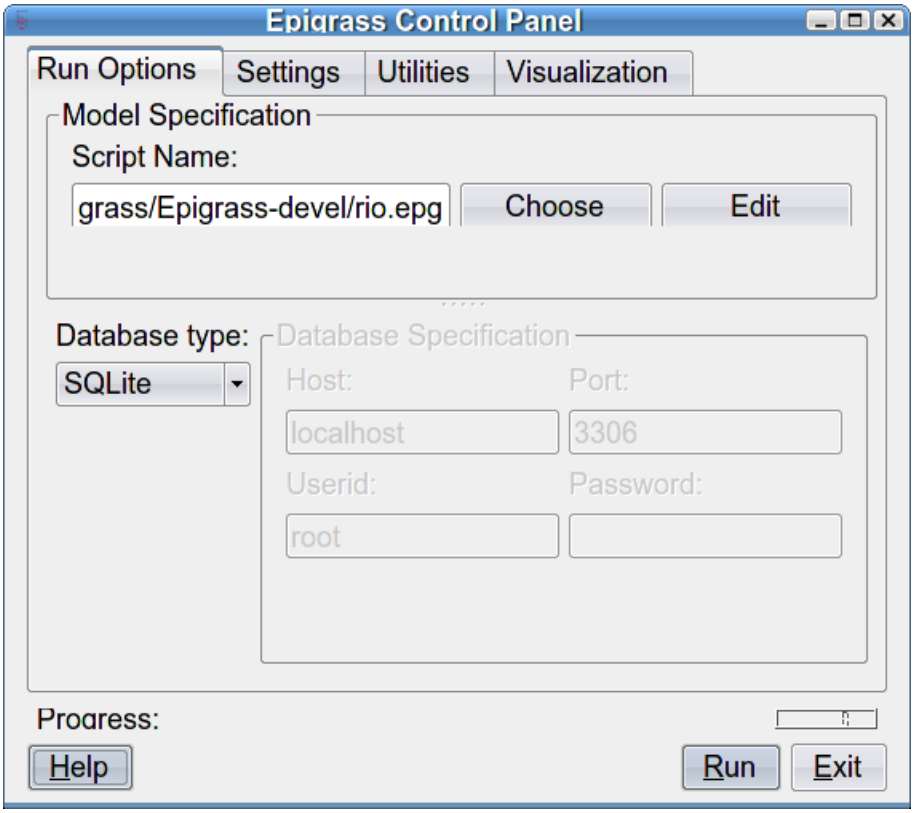


Figure 2: Epigrass graphical user interface.

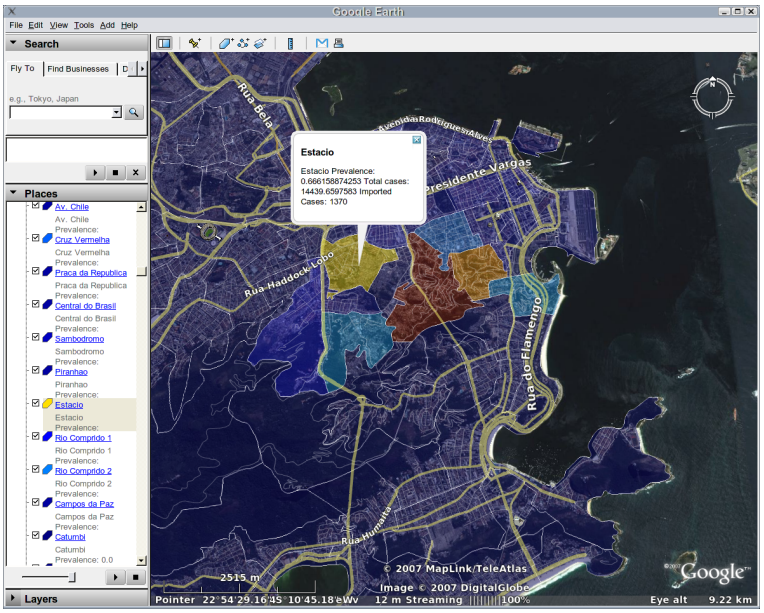


Figure 3: Epigrass output visualized on Google-Earth.

Figures

Figure 1 - Architecture of the Epigrass system.

Figure 2 - Epigrass graphical user interface.

Figure 3 - Epigrass output visualized on Google-Earth.

Figure 4 - Effect of degree(a) and betweeness(b) of entry node to the speed of the epidemic.

Figure 5 - Spread of the epidemic starting at the city of Salvador, a city with relatively small degree. The number next to the boxes indicated the day when each city developed its first autoctonous case.

Figure 6 - Cost in vaccines applied vs. benefit in cases avoided, for an epidemic starting at the highest degree city (São Paulo).

Figure 7 - Cost in vaccines applied vs. benefit in cases avoided, for an epidemic starting at a relatively low degree city(Salvador).

Figure 8 - Workflow for a typical Epigrass simulation

Figure 9 - Epigrass 3D animation output.

Tables

Table 1 - Sample table title

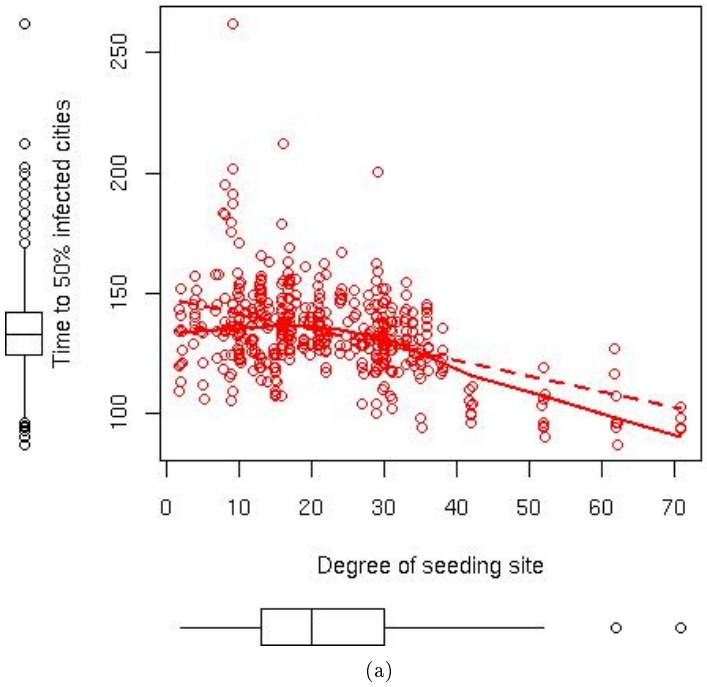
Additional Files

Additional file 1 — Sample additional file title

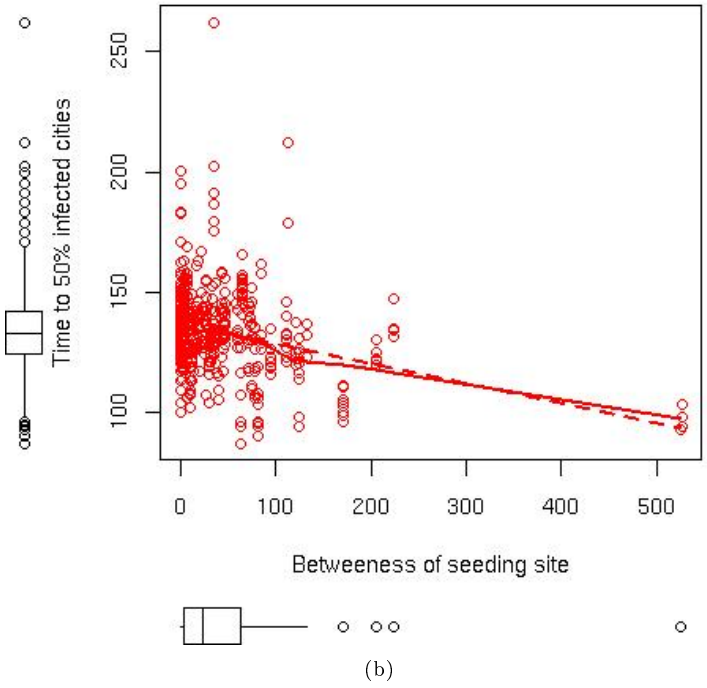
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.



(a)



(b)

Figure 4: Effect of degree(a) and betweeness(b) of entry node to the speed of the epidemic.

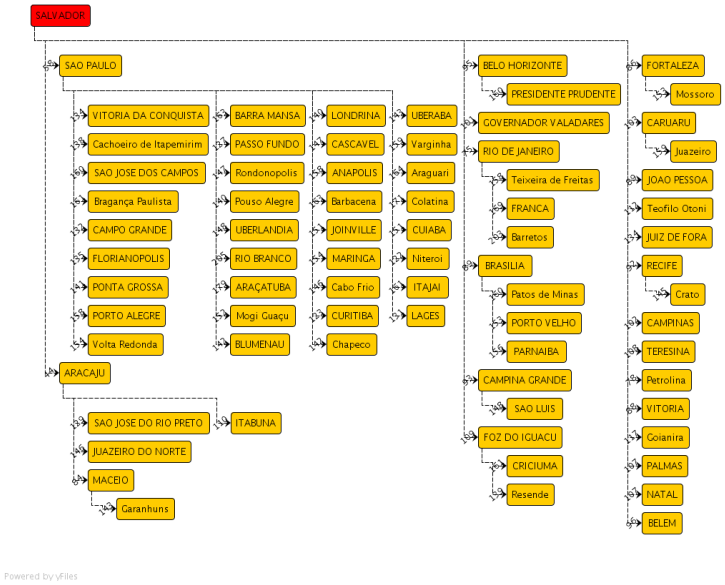


Figure 5: Spread of the epidemic starting at the city of Salvador, a city with relatively small degree. The number next to the boxes indicated the day when each city developed its first autoctonous case.

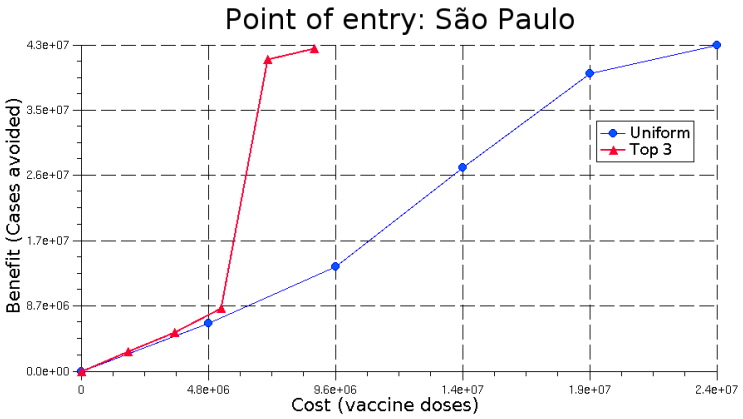


Figure 6: Cost in vaccines applied vs. benefit in cases avoided, for an epidemic starting at the highest degree city (São Paulo).

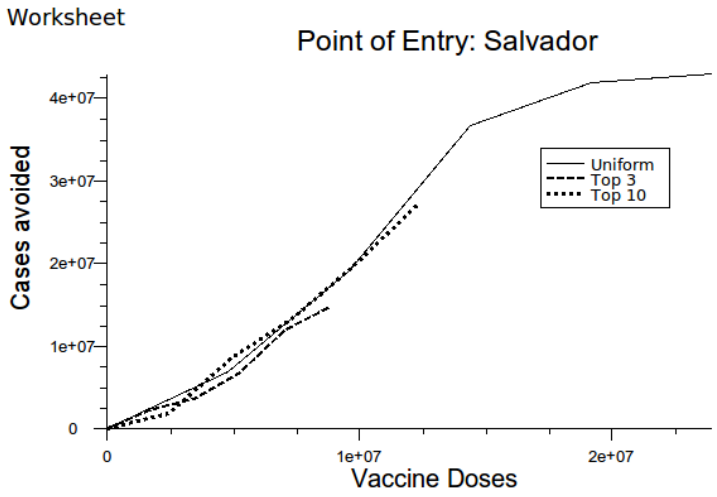


Figure 7: Cost in vaccines applied vs. benefit in cases avoided, for an epidemic starting at a relatively low degree city(Salvador).

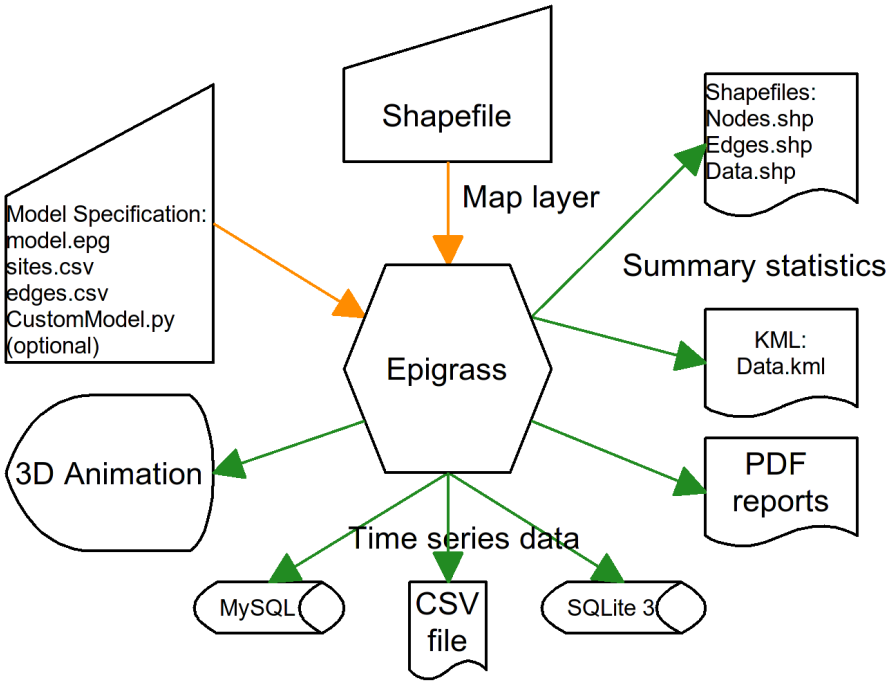


Figure 8: Workflow for a typical Epigrass simulation

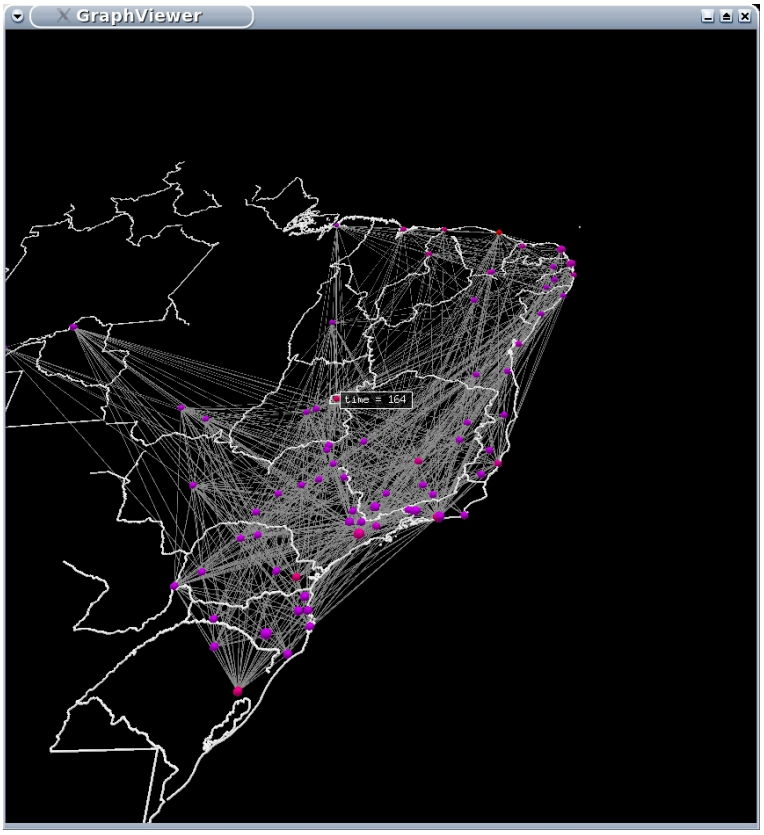


Figure 9: Epigrass 3D animation output.

Table 1: Parameters used in the models and their meaning. Parameter n and θ have their values derived stochatically during the simularion, therefore their values are not given here.

Symbol	**Meaning**.	**Value**
β	contact rate (t^{-1})	$[1.4, 2.27]$[1]
θ	number of infectious visitors	
α	mixing parameter	1
n	number of visitors	
N	population ($S + E + I + R$)	city population
r	fraction of I recovering from infection (day^{-1})	0.2
e	fraction of E becoming infectious (day^{-1})	0.2