ISDS 2016 Conference Abstracts

A Spatial Biosurveillance Synthetic Data Generator in R

Drew Levin* and Patrick Finley

Sandia National Laboratories, Albuquerque, NM, USA

Objective

To develop a spatially accurate biosurveillance synthetic data generator for the testing, evaluation, and comparison of new outbreak detection techniques.

Introduction

Development of new methods for the rapid detection of emerging disease outbreaks is a research priority in the field of biosurveillance. Because real-world data are often proprietary in nature, scientists must utilize synthetic data generation methods to evaluate new detection methodologies. Colizza et. al. have shown that epidemic spread is dependent on the airline transportation network [1], yet current data generators do not operate over network structures.

Here we present a new spatial data generator that models the spread of contagion across a network of cities connected by airline routes. The generator is developed in the R programming language and produces data compatible with the popular `surveillance' software package.

Methods

Colizza et. al. demonstrate the power-law relationships between city population, air traffic, and degree distribution [1]. We generate a transportation network as a Chung-Lu random graph [2] that preserves these scale-free relationships (Figure 1).

First, given a power-law exponent and a desired number of cities, a probability mass function (PMF) is generated that mirrors the expected degree distribution for the given power-law relationship. Values are then sampled from this PMF to generate an expected degree (number of connected cities) for each city in the network. Edges (airline connections) are added to the network probabilistically as described in [2]. Unconnected graph components are each joined to the largest component using linear preferential attachment. Finally, city sizes are calculated based on an observed three-quarter power-law scaling relationship with the sampled degree distribution.

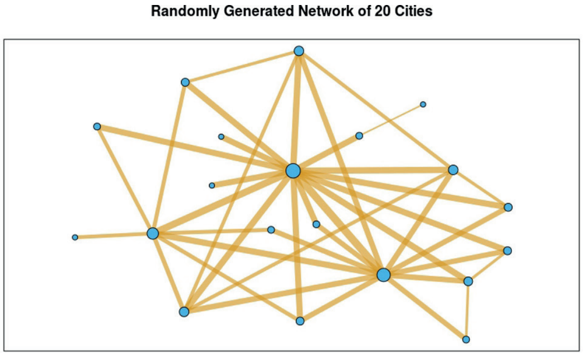
Each city is represented as a customizable stochastic compartmental SIR model. Transportation between cities is modeled similar to [2]. An infection is initialized in a single random city and infection counts are recorded in each city for a fixed period of time. A consistent fraction of the modeled infection cases are recorded as daily clinic visits. These counts are then added onto statically generated baseline data for each city to produce a full synthetic data set. Alternatively, data sets can be generated using real-world networks, such as the one maintained by the International Air Transport Association.

Results

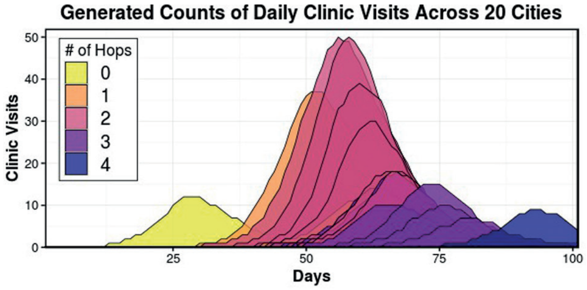
Dynamics such as the number of cities, degree distribution power-law exponent, traffic flow, and disease kinetics can be customized. In the presented example (Figure 2) the outbreak spreads over a 20 city transportation network. Infection spreads rapidly once the more populated hub cities are infected. Cities that are multiple flights away from the initially infected city are infected late in the process. The generator is capable of creating data sets of arbitrary size, length, and connectivity to better mirror a diverse set of observed network types.

Conclusions

New computational methods for outbreak detection and surveillance must be compared to established approaches. Outbreak mitigation strategies require a realistic model of human transportation behavior to best evaluate impact. These actions require test data that accurately reflect the complexity of the real-world data they would be applied to. The outbreak data generated here represents the complexity of modern transportation networks and are made to be easily integrated with established software packages to allow for rapid testing and deployment.



Randomly generated scale-free transportation network with a power-law degree exponent of λ=1.8. City and link sizes are scaled to reflect their weight.



An example of observed daily outbreak-related clinic visits across a randomly generated network of 20 cities. Each city is colored by the number of flights required to reach the city from the initial infection location. These generated counts are then added onto baseline data to create a synthetic data set for experimentation.

Keywords

Simulation; Network; Spatial; Synthetic; Data

Acknowledgments

Funding provided by Sandia National Laboratories LDRD program. Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. The role of the airline transportation network in the pre-diction and predictability of global epidemics. Proceedings of the National Academy of Sciences, 103(7):2015–2020, feb 2006.

[2] Fan Chung and Linyuan Lu. Connected Components in Random Graphs with Given Expected Degree Sequences. Annals of Combinatorics, 6(2):125–145, nov 2002

***Drew Levin**

E-mail: dlevin@sandia.gov