Open Report

Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 2.0

Network Dynamics and Simulation and Science Laboratory
Virginia Polytechnic Institute and State University
1880 Pratt Drive, Building XV
Blacksburg VA 24061
http://ndssl.vbi.vt.edu

NDSSL-TR-07-003

1 Introduction and Summary

This technical report describes data sets that are being released to the larger academic community for research. The data sets are based on detailed microscopic simulation-based modeling and integration techniques and are being released as **Data Set Version 2.0**. Data Set Version 1.0 was made available in January 2006 [7]. We expect to make available new and enhanced synthetic data products, including new cities and regions, on a regular basis. The data set provided represent a synthetic population of the city of **Portland**. The reader is referred to [2, 3, 6, 7, 11, 12] for additional details about the construction of these datasets. The data sets are released as compressed ASCII files. The following synthetic datasets are available.

1. A set of synthetic individuals in the city of Portland, USA, and a small number of demographic attributes for each synthetic individual. Different from version 1.0. Contained in the files `demographics-person-portland-1-v2.dat`and `demographics-household-portland-1-v2.dat`, see Tables 2 and 3 for examples.

2. A set of aggregated activity locations, two per roadway link in the city of Portland that have an (x, y) coordinate. Same as version 1.0, but numbered differently. Contained in the file `location-portland-1-v2.dat`. See Table 1 for an example.

3. A list of daily activities for each individual. This is different than the version 1.0 dataset. Contained in file `activities-portland-1-v2.dat`. See Table 5 for an example.

4. An instance of a time varying social contact network for a normative day, derived from the daily activities. This is a different network from version 1.0. Contained in the file `contact-portland-1-v2.dat`. See Table 7 for an example.

5. A description of disease evolution over the network discussed above when a specific set of individuals are infected. The evolution of disease is based on the **Simdemics** simulation systems and indicates when, where and from whom a person became infected and the disease state of other individuals at each location. This is based on the new social network. Contained in the files `dendrogram-<inf>-<rep>-<vac>.dat`. See Table 8 for an example.

The synthetic data is produced using **Simfrastructure**: a high-performance, service-oriented, agent-based modeling and simulation system for representing and analyzing interdependent infrastructures. Simfrastructure is used for the representation and analysis of interdependent urban infrastructures such as transportation, public health, energy, financial (commodity markets) in a service-oriented environment. Simulation environments such as Simfrastructure allow analysis of urban infrastructure interdependencies through integrated functional data flow architectures. Simfrastructure can be viewed as representing functioning virtual cities. One of its unique features is its ability to represent entire urban populations at the level of individuals, including their activities, movements and locations. The ability to generate an urban population, move each person on a second-by-second basis, and monitor the individual's interaction with other individuals, as well as with the physical infrastructure, enables greater understanding of infrastructure operations and interdependencies at an extreme but practical level of detail.

Simdemics is a group of closely related models and simulations for studying the spread of infectious disease through the urban infrastructures of Simfrastructure. Simdemics models the transmission of disease from one individual to another, the progression of the disease within an individual, and the effect of interventions on the disease spread. Interventions can be both pharmaceutical (e.g., antivirals and vaccines), and non-pharmaceutical (i.e., school closures and quarantines).

Location Id	Y Offset	X Offset
1	490844.1	5069529.5
2	490836.6	5069548.5
3	490563.0	5069540.0
4	490544.9	5069532.0
5	481259.5	5068943.0
6	481240.8	5068936.0
7	481761.4	5068519.0
8	481742.6	5068526.0
9	481572.2	5068216.0
10	481560.1	5068200.0

Table 1: Sample location file

Data set 1.0 contained only one random instance created by the **EpiSims** simulation system. Data Set 2.0 is produced using a new implementation of the models from **EpiSims** called **Simdemics** and provides the results of a more detailed experimental design.

The data sets are available from `http://ndssl.vbi.vt.edu/opendata/` for academic purposes and are released under a Creative Commons Attribution Noncommercial-ShareAlike license (`http://creativecommons.org/licenses/by-sa/2.5/`). Non-commercial use is allowed as long as the source of the data is acknowledged and any derivative data sets are released under the same license. For use in situations not covered under this license, please contact NDSSL. Any individual or organization using this data is (i) required to cite this technical report as the source of data, (ii) requested to, whenever possible, cite one of the appropriate technical papers that contain additional information about the models and methods (iii) requested to, whenever possible, send a citation and/or a copy of their work to `ndssl-data@vbi.vt.edu`.

2 Synthetic Proto-Populations and Locations

The proto-population information is contained in three files. Location file `location-portland-1-v2.dat`contains location information in form of X and Y Universal Transverse Mercator (UTM) coordinates (Zone 10) in meters (see Table 1). Information regarding synthetic individuals is stored in `demographics-person-portland-1-v2.dat`and `demographics-household-portland-1-v2.dat`(see Table 2 and Table 3).

The demographics contain, for each individual, their household id, age, gender (1-male, 2-female), worker status (1-works, 2-does not work), and relationship to the head of household (1-spouse, partner, or head of household, 2-child, 3-adult relative, 4-other). The household demographics contain household income category, number of people in the household, location and sublocation of the home, number of vehicles in the household, and number of household members who work. The household income is divided into 14 categories representing income in dollars converted to income bins as shown in Table 4.

3 Activities of the Synthetic Proto-Population

The file `activities-portland-1-v2.dat` describes the daily activities of each synthetic individual (see Table 5 and 6). Each location is divided into a number of sublocations. These sublocations can be thought of as separate rooms within the location.

Person ID	Household ID	Age	Gender	Worker	Relationship
2509159	2201175	42	1	1	1
2509160	2201175	43	2	1	1
2509161	2201175	17	1	2	2
2509162	2201176	41	1	1	1
2509163	2201176	11	1	2	2

Table 2: Sample person demographics files

Household ID	Income	Size	Location Id	Sublocation Id	Vehicles	Workers
2201175	13	3	53571	1004	3	2
2201176	11	2	53834	1000	3	1

Table 3: Sample household demographics files

HHIncome Value	Household Income range in $
1	$0 - 4999$
2	$5000 - 9999$
3	$10000 - 14999$
4	$15000 - 19999$
5	$20000 - 24999$
6	$25000 - 29999$
7	$30000 - 34999$
8	$35000 - 39999$
9	$40000 - 44999$
10	$45000 - 49999$
11	$50,000 - 54,999$
12	$55,000 - 59,000$
13	≥ 60000
14	< 0

Table 4: Table showing the income groups used to divide individuals.

Household Id	Person Id	Activity Number	Activity Type	Start Time	Duration	Location Id	Sublocation Id
2201175	2509159	1	0	1	25199	53571	1004
2201175	2509159	2	1	25200	30600	48892	2003
2201175	2509159	3	0	57600	28800	53571	1004
2201175	2509160	1	0	1	28918	53571	1004
2201175	2509160	2	8	28919	13679	52506	6000
2201175	2509160	3	1	43200	23400	52563	2035
2201175	2509160	4	4	67500	11700	49235	4005
2201175	2509160	5	0	80100	6300	53571	1004
2201175	2509161	1	0	1	29699	53571	1004
2201175	2509161	2	8	29700	24300	53244	6001
2201175	2509161	3	0	54900	900	53571	1004
2201175	2509161	4	1	57600	28800	53429	2000

Table 5: Sample activity file

Activity Type	Activity Description
0	Home
1	Work
2	Shop
3	Visit
4	Social/Recreational
5	Other
6	Serve Passenger
7	School
8	College

Table 6: Table showing the types of activities used in the data sets.

Person A Id	Activity Type	Person B Id	Activity Type	Contact Duration (seconds)
2000020	0	2000019	0	86399
2000040	0	2000038	0	40499
2000040	0	2000039	0	48599
2000060	0	2000059	0	43497
2000060	1	2000009	1	2099
2000060	1	2000059	1	11099

Table 7: Sample contact network file.

4 Social Contact Network

A social contact network captures the interaction between individuals moving through an urban region [8–10]. The file `contact-portland-1-v2.dat` (see Table 7) is an undirected edge list representation of the static people-to-people contact graph, obtained using the Simdemics sublocation model. Two individuals have an edge if and only if they are ever co-located at the same location and sublocation in the activity file.

Each line corresponds to an edge, including the ID's of two people involved, their activity types at the time of this contact, and the duration of the contact in seconds. The edges in this file are bi-directional, but in order to save space only one direction is represented in the file.

5 Output of Disease Dynamics on the Synthetic Social Network

We have used **Epi-Fast**, a part of **Simdemics**, a fast stochastic simulation system to generate disease dynamics for the social network described in Section 4. Each run of **Epi-Fast** over a given social network, a random initial condition, and exogenous interventions if there are any, generates a random instance of disease dynamics. The Spatial-Temporal propagation of the disease in such a random realization is described by a *dendrogram*.

The dendrogram contains the infected people, and for each infected individual, it records his ID, the day he gets infected, the location and sublocation at which he gets infected, a generation number, and the ID of the individual (called *parent*) that infects him. The generation of an infected individual is one more than that of his *parent*; the initially infected individuals have a generation number of 1 and parent of -1. See Table 8.

5.1 Details of experimental design used to generate dendrograms

It is often interesting to see how the disease propagation would have changed had there been interventions. To this end, we have simulated the epidemic trajectory with different subsets of people vaccinated. The

Infected Id	Day of Infection	Location of Infection	Sublocation	Generation	Parent Id
2000001	99	9984	6001	23	2036889
2000002	101	5212	5014	27	2016388
2000003	104	32133	2010	24	2263525
2000005	105	5212	5012	26	2019218
2000006	113	5212	5034	28	2017265
2000011	102	31640	2000	24	3189333
2000012	151	9985	6000	36	2000034
2000016	96	9982	6000	21	2181004
2000017	119	5212	5015	27	2152276

Table 8: Sample dendrogram.

experimental design considers three independent variables: (i) disease property (3 levels) (ii) the method by which individuals are picked to be vaccinated (2 levels) and (iii) the number of individuals that are picked (2 levels). The various levels for each treatment are described below.

We have studied 3 cases for the disease property setting: *low, medium* and *high* infectivity. Intuitively the disease property setting tells us the chance that a given individual infects his neighbor. In case of low infectivity, there may or may not be a large epidemic outbreak. In case of high infectivity, the disease almost surely spreads to at least 80% of the whole population. The case of medium infectivity is in between these two extreme cases.

The data set provided gives the result of vaccinating a subset of individuals. Other forms of intervention can also be studied and data for these interventions will provided in future versions.

1. **Case 1:** is the Base case in which we have no interventions.

2. **Case 2:** A small portion (about 5%) of the population are chosen randomly to be vaccinated; the chosen ID's are in file `Random.Vaccination.5percent.txt`.

3. **Case 3:** a large portion (about 30%) of the population are chosen randomly to be vaccinated; the chosen ID's are in file `Random.Vaccination.30percent.txt`.

4. **Case 4:** the top 5% of the population that have largest topological degrees (neighbors) in the social network are vaccinated; their ID's are in file `High.Degree.Vaccination.5percent.txt`.

5. **Case 5:** the top 30% of the population that have largest topological degrees in the social network are vaccinated; their ID's are in file `High.Degree.Vaccination.30percent.txt`.

In other words, cases 2 and 3 provide data when individuals to be vaccinated are picked randomly; cases 4 and 5 deal with the case when individuals are picked based on how many other individuals they meet.

A total of 150 dendrograms have been created (3 levels of infection, 10 replicates, 1 initial condition and 5 different intervention strategies including the case of no intervention). Four individuals are initially infected and the base case and the interventions are simulated with this as the initial condition. A different set of individuals is infected in each of the 10 replicates, for each infectivity (a total of 30 sets of infected individuals).

They are stored in files of the form `dendrogram-<inf>-<rep>-<vac>.dat`, where `<inf>` is one of *low*, *med*, *high*, `<rep>` is an integer in the range 0 through 9, and `<vac>` is one of *none*, *rand5*, *rand30*, *degree5*, *degree30*.

`<inf>` represents the infectivity of the individuals in the population, `<rep>` the replicate number (each case has been run 10 times), and `<vac>` represents the part of the population that has been vaccinated.

References

[1] K. Atkins, C. Barrett, C. Homan, A. Marathe, M. Marathe and S. Thite. "Agent Based Economic Analysis of Deregulated Electricity Markets", *6th IAEE European Conference*, Zurich, Switzerland, September 2004.

[2] C. Barrett, R. Beckman, K. Berkbigler, K. Bisset, B. Bush, K. Campbell, S. Eubank, K. Henson, J. Hurford, D. Kubicek, M. Marathe, P. Romero, J. Smith, L. Smith, P. Speckman, P. Stretz, G. Thayer, E. Eeckhout, and M.D. Williams. TRANSIMS: Transportation Analysis Simulation System. Technical Report LA-UR-00-1725, Los Alamos National Laboratory Unclassified Report, 2001. An earlier version appears as a 7 part technical report series LA-UR-99-1658 and LA-UR-99-2574 to LA-UR-99-2580.

[3] R J Beckman et al, TRANSIMS-Release 1.0: The Dallas-Fort Worth Case Study, Technical Report LA-UR-97-4502, Los Alamos National Laboratory, 1997.

[4] C. Barrett, S. Eubank, V. Anil Kumar, M. Marathe. Understanding Large Scale Social and Infrastructure Networks: A Simulation Based Approach, in *SIAM news*, March 2004. Appears as part of Math Awareness Month on The Mathematics of Networks.

[5] Barrett C, Marathe M, Smith J, Ravi S, (2002) A mobility and traffic generation framework for modeling and simulating ad hoc communication networks. ACM Symposium on Applied Computing (SAC), pp. 122-126

[6] R. J. Beckman, K. A. Baggerly, and M. D. McKay, Creating synthetic base-line populations, *Transportation Research Part A – Policy and Practice* 30 (1996) 415–429.

[7] Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 1.0, Technical Report NDSSL-TR-06-006, Network Dynamics and Simulation and Science Laboratory, Virginia Tech, 2006. `http://ndssl.vbi.vt.edu/Publications/ndssl-tr-06-006.pdf`

[8] S. Eubank, H. Guclu, V.S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai and N. Wang,.Modeling Disease Outbreaks in Realistic Urban Social Networks, Nature, 429, pp. 180-184, May (2004).

[9] S. Eubank, V.S. Anil Kumar, M. Marathe, A. Srinivasan and N. Wang. Structural and Algorithmic Aspects of Large Social Networks, Proc. 15th ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 711-720, (2004).

[10] S. Eubank, V.S. Anil Kumar, M. Marathe, A. Srinivasan and N. Wang. Structure of Social Contact Networks and their Impact on Epidemics. to appear in AMS-DIMACS Special Volume on Epidemiology, (2005).

[11] Speckman P, Vaughn K, Pas E (1997) Generating Household Activity-Travel Patterns (HATPs) for Synthetic Populations. Transportation Research Board 1997 Annual Meeting.

[12] Speckman P, Vaughn K, Pas E (1997) A Continuous Spatial Interaction Model: Application to Home-Work Travel in Portland, Oregon. Transportation Research Board 1997 Annual Meeting