Multi-Coaffiliation Networks and Public Health Applications

O. Loza, I. Gomez-Lopez
Computational Epidemiology Research Lab
University of North Texas
Denton TX, USA
Email: {oloza, igomez}@unt.edu

A. R. Mikler
Center of Computational Epidemiology and Response Analysis
University of North Texas
Denton TX, USA
Email: mikler@unt.edu

***Abstract*—Infectious diseases are a global concern. The challenge for public health bodies relies upon optimizing the distribution of scarce or costly control measures to maximize their impact on the outbreak dynamics. Risk identification has focused on schools and child-care centers mainly because they represent dense masses of highly immunologically naive hosts for the pathogens. To advance the design of mitigation strategies, epidemiology researchers have broaden their perspective through the use of computational tools designed to provide decision support for multiple scenarios.**

To identify at-risk populations, we propose a computational algorithm that recreates a realistic social model of the school system of a selected study place. It is a known fact that childhood diseases are spread through the social contacts that occur in the classrooms while schools are in session. Through synthetic reconstruction, the algorithm generates a synthesized population database. The demographic simulations are created at the level of individuals, households and schools. Then a school to school network is built as a representation of the social model. The algorithm outputs a new graph B', representing the Multi- Coaffiliation Network (MCN) with number of vertices of order $O(S)$, where S represents the number of schools. The resulting weighted network includes a value associated with each school as a possible intervention location. The risk-evaluation of the schools in the network can be derived in a wide range of applications in both research and public policy analysis.

***Keywords*-Computational Epidemiology; Algorithms; Affiliation Networks**

I. Introduction

Disease spreading models improve implementation of disease control and enhance public health surveillance. The analysis of epidemic outbreaks is crucial and is mainly based of information reported after an event has occurred. How to measure and target intervention strategies is a complex problem. Research on novel infections showed that targeting intervention measures is more effective if the group with the highest risk can be identified [1]. Disease control measures comprehend vaccine, antiviral treatment and prophylaxis and non-pharmaceutical interventions such quarantine, isolation, school closure, and social distancing [2]. In this regard, timing is a key factor for interventions, scheduled vaccination are part of almost any nation public health policy, however social distancing occurs once cases of a particular disease have been identified. For many infectious diseases, vaccination is the most effective means of control and their objective is to immunize a sufficiently high proportion of the individuals in the community to prevent an epidemic. It is believed that interventions targeted at school-aged children, should be most effective in the early stages of an outbreak[3]. In particular, if the incidence becomes comparable among children and adults, measures will be significantly more valuable at the start of the pandemic than they would have been later on[1]. For instance, after a transversal comparison of several interventions for influenza, it was discovered that early detection and initiation of measures and school closure play important roles in reducing influenza transmission[2]. Disease outbreaks occur due to the existence of a wider network that individuals are not well aware of. Through the identification of communities and community bridges in the social model, targeted immunizations could be more effective[4]. This concept has also been studied at the granularity of households and transmission within and between households the household for isolation, quarantine, vaccination or prophylactic treatment [5].

Lowering the epidemic severity through reducing school-age contacts is an important component of the U.S. and other countries mitigation strategy [6]. The use of meta-population models and virtual populations has been extensive during the past years for the exploration on theoretical models applied to epidemiology [7]. These models been used for risk assessment as well, designed to study different communicable [8] and non-communicable [7] diseases. In these models, synthetic populations based on the information obtained from the public census are used to run simulated scenarios. Intervention strategies can then be compared and contrasted based on their impact on the disease dynamics. The simulations allow to identify those parts of model that have the most effect on the dynamic of the disease [9].

Determining the mixing network in full extend, requires complete knowledge of every individual in a population. In addition, complete recall of the person's relationships are needed to construct the network. These tasks seem unattainable and impractical even for small groups. To establish a workable framework, researchers initially applied known datasets from social networks based on random graphs

DOI: 10.5176/2251-3140_2.1.32

with arbitrary degree distribution that have exactly solvable models[10], random graphs with tunable clustering [11] and stochastic process with global [12] and local contacts[13], [14], [15].

We propose a model that focuses on the school system of an artificially reconstructed population. Synthetic populations were initially created to estimate traffic needs and development [16] but were quickly adopted by epidemiology researches to demonstrate a strong correlation between local demographic characteristics and pandemic severity [17]. Microsimulation models are snapshots of the entire population of a selected location and are utilized mainly in Agent-Based models. The contact networks produced after the complete recording of the agents contacts are normally of the order of the total population N. The principal input for the proposed algorithm is the synthetic population dataset containing people and households records. Additionally, information of public schools enrollment and location is needed. The outcome of the algorithm is a weighted graph that represents schools and weighted links among schools. The size of the resulting network is logarithmically smaller that the initial contact network and it focuses exclusively on the relevance of each school in the educational system.

II. Methods

A. Model Background and Related Work

The algorithm is simple and applicable to any region where population census and public schools information are available. Synthetic reconstruction is a process that creates data records describing socio-demographic features of households (HH) and households members (P) residing in the study area. The general generation process requires an *aggregate dataset* that contains the marginal distribution of the variables estimated for the year of study and a *disaggregate dataset* or a sample of records with complete information about Ps and HHs in the population. Given the aggregate and disaggregate datasets, the population records are produced by selecting sample records from the disaggregate dataset to meet the marginal distribution given by the aggregate dataset [18]. The process outputs data records with the selected demographic descriptors updated to the correspondent year of study. The methodology has been applied to diverse sources of information for different geographies:the city of Portland, Oregon [19], New York [20], 50 states and the district of Columbia [21], Switzerland [22], Belgium [23], Singapure [24]. Comparisons among methodologies has also been done by [25], and [26]. The methodology presented in [16] and [18] requires US Census Bureau Summary Files (SFs) [27] and Public Use Micro data Sample (PUMS). PUMS Files for the 2010 Census are scheduled for release on December 2012 through April 2013. The census 2000 was the latest complete information available at the moment this research was been developed. Tables I and II contain a list of control variables used for the synthetic reconstruction of Ps, HHs, and schools (Ss).

B. Selected Variables

Table I
Variables and correspondent Columns in SF1

Control	Description	SF1
P-Age	Age of the individual	P12
P-Race	Race of the individual	P7
P-Gender	Gender of the individual	P12
HH-Fam	Family o non-family household	P26
HH-Size	Size of the household	P26
HH-Type	Type of the household	P20
HH-Child	Households by presence of people under 18 years old	P19

Table II
Schools Information Used

Control	Description
S-Type	School Type
S-Enrollment	Student Enrollment of School
S-LocationX	Location of the School Longitud component
S-LocationY	Location of the School Latitud component
S-HighLevel	Highest level of the School
S-LowLever	Lowest level of the School

While the definition of "community" is domain dependable, one can take the notion of a collection of similar entities that interact unusually frequently. In a social context, families are the smallest representation of community. As groups, communities have a dynamic behavior similar to the entities they are composed by. The identification of communities can be applied to multiple problems [28]. Different computational frameworks have been proposed [29] to study communities and their influence on disease dynamics. In order to study the school system, we considered each school as a community of households. Additionally, to construct links between schools, the number of households that the schools have in common was calculated.

C. Definition of Multi-Coaffiliation Networks (MCNs)

The *School Affiliation Network Discovery (SAND)* Algorithm constructs the Multi Coaffiliaiton Networks (MCN) of

the study area and is defined as follows. A simple undirected bipartite graph $B = (V, A)$, $V = (S \cup HH)$ is constructed to represent the affiliation network formed by schools S and households HH, where an edge $(hh_i, s_j) \in A$ represents a child belonging to household $hh_i \in H$ attending school $s_j \in S$.

The function $\mathcal{A}(hh_i, s_j)$ is a binary function that identifies the membership of a hh_i to s_j.

$$\mathcal{A}(hh_i, s_j) = \begin{cases} 1 & \text{if } (hh_i, s_j) \in A \\ 0 & \text{otherwise} \end{cases}$$

The way $\mathcal{A}$ is defined may vary from region to region. In the US, the School Districts (SDs) and in Texas the Independent School Districts (ISDs) are responsible for delimiting the School Attendance Zones (SAZs) and making the information available to the public. Maps can be publicly accessed in diverse formats, regularly through SDs websites. The attendance zones may also be updated as the population changes and new establishments are created. Although there is not a defined standard for how the information is presented to the public, in the US, it can be obtained in PDF format, interactive maps or a listing of streets. Nevertheless, lack of information in other geographic regions may be a big concern. For other geographies the selection process for attending public schools may be mainly driven by density of the population, religion preference or acquisitive potential. The following assumptions are made about people attending school:

- The definition for distance is Euclidean distance.
- The closest available and matching type school is chosen.
- If the household has more than one member attending a particular school type (i.e. elementary school) then all the members of the household are assigned to the same school.

To generalize the function $\mathcal{A}$ a modifier $D(\mathcal{A})$ can be applied. For example, the function D, may be Euclidean distance (Equation 1) in order to draw the links between schools and HHs with school-age children.

$$D = \sqrt{(x(h_i) - x(s_j))^2 + (y(h_i) - y(s_j))^2} \qquad (1)$$

In Equation 1 the functions $x(p)$, and $y(p)$ stand for the latitude and longitude of a given point p.
The scheme will yield a Voronoi tessellation for each type of school in the school system, at the study area. The accuracy could potentially be estimated by comparing known SAZs maps and the Voronoi diagram. A comparison based on Denton ISD is shown in Figure 1. The distance function D can also be defined as Manhattan Distance, Minkowski distance or other and evaluated to reduce the approximation error.

B is then transformed into a weighted graph $B' = (S, EE)$ or (MCN), $S = (E \cup M \cup H)$. where:

$$e_k = (s_i, s_j) \in EE = \forall_{hh_k \in HH} \alpha(hh_k, s_i) \times \mathcal{A}(hh_k, s_j) \qquad (2)$$

And the function $W(e_k) \geq 0$ representing the weight of the edge e_k is defined as

$$W(e_k) = \sum_{hh_k \in HH} \{(hh_k, s_i), (hh_k, s_j)\} \in A \qquad (3)$$

The construction of B' is schematized in Figure 2

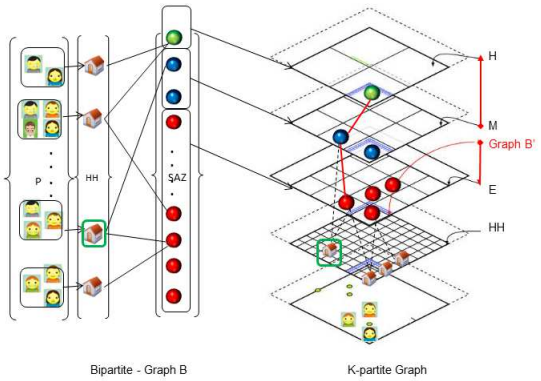

Figure 2. Construction of Graph B'

D. Algorithms

Select a geographical unit $\mathcal{C}$. $\mathcal{C}$=$\{HH, P, S\}$ Where:
$HH = \{hh_1, h_2, \ldots, hh_i, \ldots, hh_n\}$ is the set of households h_i that live in a house located insid the area of $\mathcal{C}$.
$P = \{p_{(i,1)}, p_{(i,2)}, p_{(i,3)}, \ldots\}$ represent that person $p_{(i,j)}$ who belongs to household i. For instance, the largest household size registered in Texas at five percent sample file SF1, is thirteen. Households of size one are not considered.
Each person also represents a set of attributes:
$p_{(i,j)} = \{[age], [gender], [attends\_school], [location], ...\}$
The set of characteristics come from the description on the file SF1, census 2000 and listed in Table I.
$S = \{s_1, s_2, \ldots, s_k, \ldots, s_m\}$, where school s_k, is belongs to an SD located inside $\mathcal{C}$
Each school also has a set of attributes, listed in Table II:
$s_k = \{[lower\_level], [higher\_level], [location], ...\}$
The algorithm used to link schools is as follows:
Algorithm 1 runs in $O(S \times P)$ to account for all the affiliations household-school. Algorithm 2 in the worse case scenario runs in $O(S^2)$.

III. Results

A. Application Example

An example of the general output the algorithm is shown in Figure 3. The left image depicts location of households,

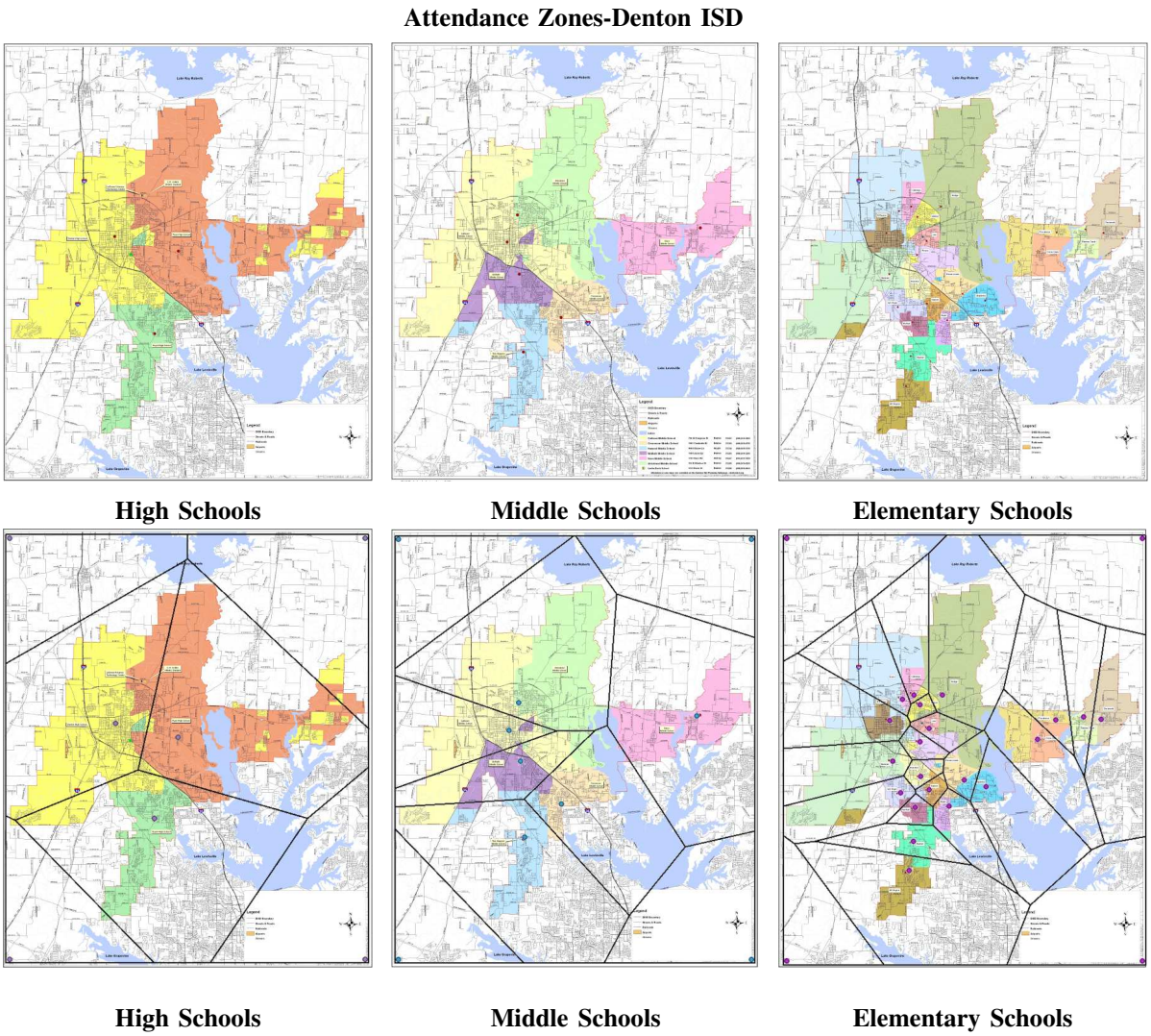

Figure 1. Voronoi Tessellation of Denton ISD Attendance Zones

Algorithm 1 Construction of School Affiliations : Graph B

Require: HH, P, S
  for $i = 1$ TO $|HH|$ **do**
    for $j = 1$ TO $hh_i[num\_people\_in\_HH]$ **do**
      if $p_{(i,j)}.attends\_schl$ is TRUE **then**
        for $k = 1$ TO $|S|$ **do**
          if $p_{(i,j)}.level \geq s_k.lower\_level$ AND $p_{(i,j)}.level \leq s_k.upper\_level$ **then**
            $psbl\_schls_{(i,k)} \leftarrow dist(hh_i.loc, s_k.loc)$
          end if
        end for
        $sschl \leftarrow min(psbl\_schls)$
        $lnk\_schlsHH(i, sschl) \leftarrow crt\_lnk(i, sschl)$
      end if
    end for
  end for

Algorithm 2 Construction of MCNs: Graph B'

Require: $S, lnk\_schlsHH$
  for $k = 1$ TO $|S|$ **do**
    $list\_HH_{[k]} \leftarrow select(k, lnk\_schlsHH)$
    $lnk\_schls_{[k]} \leftarrow select(list\_HH_{[k]}, lnk\_schlsHH)$
    for $i = 1$ **to** $|lnk\_schls_{[k]}|$ **do**
      if $(lnk\_schls[i] \neq s[k])$ **then**
        $lnk\_schls\_schls \leftarrow crt\_lnk[(k, lnk\_schls[i])]$
      end if
    end for
  end for

schools and their affiliation. The image on the right exposes the resulting school network. The output of the SAND algorithm is a graph weighted on the vertices and edges.

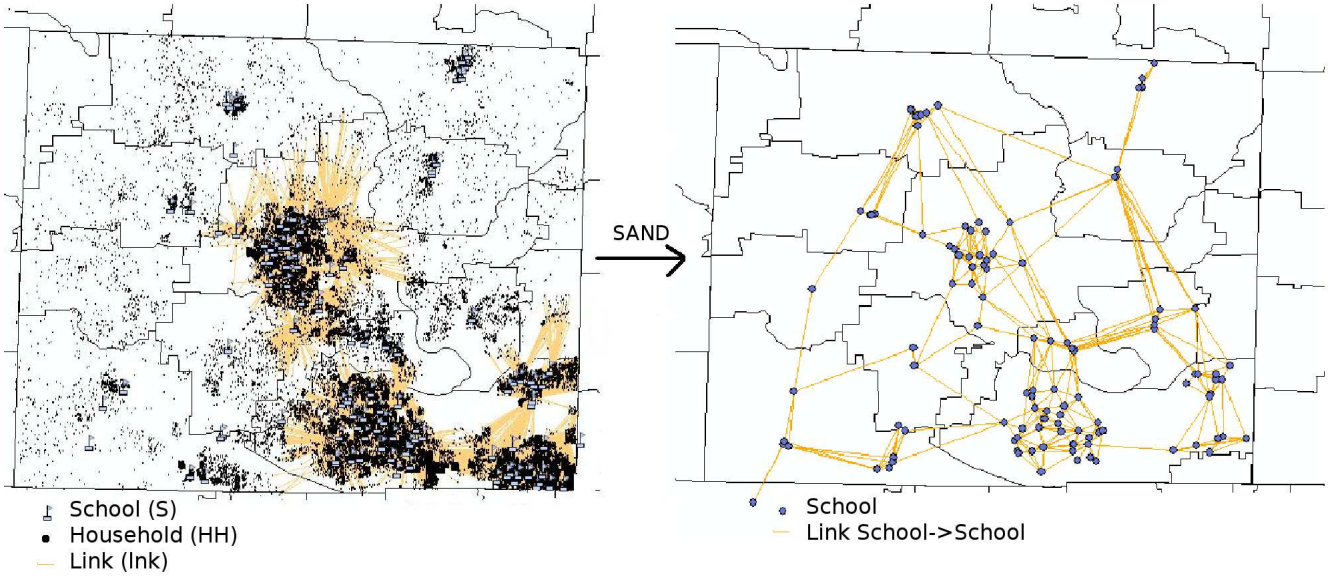

Figure 3. A example of the conversion for multiple SDs - Texas

The weight of the vertices can be considered as an intra-community measure, whereas the weight of the outgoing edges could potentially be considered as an inter-community measure. The reduction in size of the social model for the state of Texas, Denton ISD is shown in Table III.

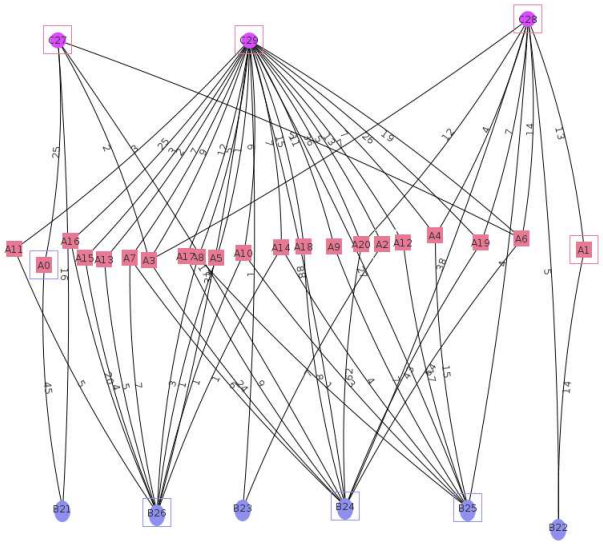

Figure 4. Denton ISD, school type :A(Elementary), B(Middle), and C(High)

Table III
TEXAS NUMBERS, DATA SOURCE: US CENSUS BUREAU AND TEXAS PUBLIC SCHOOL DIRECTORY

Variable	Description	Value
P	Total Population	25,145,561
HH	Number of Households	8,922,933
S	Number of Public Schools	8,317

Denton county and Denton ISD are shown in Table IV and Figure 4 .

IV. DISCUSSION

V. CONCLUSION AND FUTURE WORK

We present the SAND algorithm as an alternative to reduce the size and complexity of a contact network model. Due to the fact that the network represent the school system, mitigation strategies oriented to childhood diseases can be isolated and studied.

REFERENCES

[1] J. Wallinga, M. van Boven, and M. Lipsitch, "Optimizing infectious disease interventions during an emerging epidemic," *Proceedings of the National Academy of Sciences*, vol. 107, no. 2, pp. 923–928, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1073/pnas.0908491107

[2] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, D. A. T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, D. Wagener, R. Beckman, K. Kadau, C. Barrett, C. A. Macken, D. S. Burke, and P. Cooley,

Table IV
DENTON STATISTICS, DATA SOURCE: US CENSUS BUREAU AND TEXAS PUBLIC SCHOOL DIRECTORY

Variable	Description	Value
P	Total Population	662,614
HH	Number of Households	240,289
S	Number of Public Schools	30
	Total Students	22,825
S	Number of Nodes Graph B'	30
EE	Maximum Number of Edges Graph B'	73
	Is Graph B' Connected?	True

"Modeling targeted layered containment of an influenza pandemic in the united states." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 12, pp. 4639–44, Mar. 2008. [Online]. Available: http://www.pnas.org/cgi/content/abstract/105/12/4639

[3] N. Becker, "Optimal vaccination strategies for a community of households," *Mathematical Biosciences*, vol. 139, no. 2, pp. 117–132, Jan. 1997. [Online]. Available: http://dx.doi.org/10.1016/S0025-5564(96)00139-3

[4] M. Salathé and J. H. Jones, "Dynamics and control of diseases in networks with community structure." *PLoS computational biology*, vol. 6, no. 4, p. e1000736, Jan. 2010. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1000736

[5] C. Fraser, M. Research, C. Centre, O. Analysis, I. D. Epidemiology, U. Kingdom, E. Individual, H. R. Numbers, E. Epidemic, and P. One, "Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic," *America*, no. 8, 2007.

[6] H. Lempel, R. A. Hammond, and J. M. Epstein, "Center on social and economic dynamics working paper no.55," *Health Care*, no. 55, 2009.

[7] R. M. Leu M, Czene K, "Population lab: The creation of virtual populations for genetic epidemiology research. epidemiology," *Epidemiology*, vol. 18, pp. 433–440, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17486019?dopt=Abstract

[8] V. Colizza, A. Barrat, M. Barthelemy, A.-j. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions," *PLoS Medicine*, vol. 4, no. 1, 2007.

[9] D. Mollison, "The Structure of Epidemic Models," *Analysis*, no. Section 3, pp. 17–33, 1995.

[10] M. E. J. Newman, D. J. Watts, S. H. Strogatz, and E. Data, "Random graph models of social networks," *Fortune*, vol. 99, 2002. [Online]. Available: http://www.pnas.org/content/99/suppl.1/2566.full

[11] T. Britton, M. Deijfen, A. N. Lager, and M. Lindholm, "Epidemics on random graphs with tunable clustering," no. 1980, 2007.

[12] J. M. Epstein, D. M. Goedecke, F. Yu, R. J. Morris, D. K. Wagener, and G. V. Bobashev, "Controlling pandemic flu: The value of international air travel restrictions," *PLoS ONE*, vol. 2, no. 5, p. e401, 05 2007. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0000401

[13] F. Ball and P. Neal, "Network epidemic models with two levels of mixing." *Mathematical biosciences*, vol. 212, no. 1, pp. 69–87, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18280521

[14] S. Venkatachalam and A. Mikler, "Modeling infectious diseases using global stochastic field simulation," in *Granular Computing, 2006 IEEE International Conference on*, may 2006, pp. 750 – 753.

[15] A. Mikler, A. Bravo-Salgado, and C. Corley, "Global stochastic contact modeling of infectious diseases," in *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS '09. International Joint Conference on*, aug. 2009, pp. 327 –330.

[16] R. J. Beckman, K. A. Baggerly, and M. D. McKay, "Creating Synthetic Baseline Populations," *Transportation Research Part A: Policy and Practice*, vol. 30, pp. 415–429, Nov. 1996.

[17] P. Stroud, S. Del Valle, S. Sydoriak, J. Riese, and S. Mniszewski, "Spatial dynamics of pandemic influenza in a massive artificial society," *Journal of Artificial Societies and Social Simulation*, vol. 10, no. 4, p. 9, 2007. [Online]. Available: http://jasss.soc.surrey.ac.uk/10/4/9.html

[18] A. R. Pinjari, N. Eluru, R. B. Copperman, I. N. Sener, J. Y. Guo, S. Srinivasan, and C. R. Bhat, *Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP models, framework, software architecture and application results*, ser. Research Report. 40808, Texas Department of Transportation, Department of Civil, Architectural and Environmental Engineering, University of Texas Austin, Austin, 2006.

[19] S. Eubank, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, and N. Wang, "Structural and algorithmic aspects of massive social networks," in *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, ser. SODA '04. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2004, pp. 718–727. [Online]. Available: http://dl.acm.org/citation.cfm?id=982792.982902

[20] G. Martín, M.-C. Marinescu, D. E. Singh, and J. Carretero, "Leveraging social networks for understanding the evolution of epidemics," *BMC Systems Biology*, vol. 5, no. Suppl 3, p. S14, 2011. [Online]. Available: http://www.biomedcentral.com/1752-0509/5/S3/S14

[21] W. Wheaton, J. C. Cajka, B. M. Chasteen, D. K. Wagene, P. Cooley, and Ganapathi, "Synthesized Population Databases: A US Geospatial Database for Agent-Based Models," *Database*, no. May, 2009.

[22] M. Frick and K. Axhausen, "Generating synthetic populations using ipf and monte carlo techniques," *4th Swiss Transport Research Conference*, 2004. [Online]. Available: http://www.matsim.org/node/99

[23] C. C., C. E., and T. P., "Model of weekly working participation for a belgian synthetic population," *Proceedings of the European Transport. Conference (ETC) 2007*, Oct. 2007.

[24] T. Zhang, S. H. Soh, X. Fu, K. K. Lee, L. Wong, S. Ma, G. Xiao, and C. K. Kwoh, "Hpcgen a fast generator of contact networks of large urban cities for epidemiological studies," in *Proceedings of the 2009 International Conference on Computational Intelligence, Modelling and Simulation*, ser. CSSIM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 198–203. [Online]. Available: http://dx.doi.org/10.1109/CSSim.2009.46

[25] K. Müller, K. Axhausen, K. Axhausen, and K. Axhausen, *Population synthesis for microsimulation: state of the art*. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau (IVT), 2010. [Online]. Available: http://books.google.com/books?id=OTzcXwAACAAJ

[26] M. T., B. M., and D. S., "An evaluation of synthetic household populations for census collection districts created using spatial microsimulation techniques," *National Center for Social and Economic Modelling, University of Canberra, Australia*, 2002.

[27] U. S. C. Bureau. (2012, Feb.) Summary files. [Online]. Available: {http://www.census.gov/}

[28] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, p. 717, 2007. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1281192.1281269

[29] C. D. Corley and A. R. Mikler, "A computational framework to study public health epidemiology," *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pp. 360–363, 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5260636