

## Georgia State University ScholarWorks @ Georgia State University

Public Health Faculty Publications

School of Public Health

2015

# NOTES2: Networks-of-Traces for Epidemic Spread Simulations

Sicong Liu

*Arizona State University, [s.liu@asu.edu](mailto:s.liu@asu.edu)*

Yash Garg

*Arizona State, [yash.garg@asu.edu](mailto:yash.garg@asu.edu)*

K. Selçuk Candan

*Arizona State University, [candan@asu.edu](mailto:candan@asu.edu)*

Maria Luisa Sapino

*University of Torino, [marialuisa.sapino@unito.it](mailto:marialuisa.sapino@unito.it)*

Gerardo Chowell

*Georgia State University, [gchowell@gsu.edu](mailto:gchowell@gsu.edu)*

Follow this and additional works at: [https://scholarworks.gsu.edu/iph\\_facpub](https://scholarworks.gsu.edu/iph_facpub)



Part of the [Public Health Commons](#)

### Recommended Citation

S. Liu, Y. Garg, K. S. Candan, M. L. Sapino, G. Chowell. NOTES2: Networks- of-Traces for Epidemic Spread Simulations. Computational Sustainability: Papers from the 2015 AAAI Workshop.

This Conference Proceeding is brought to you for free and open access by the School of Public Health at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Public Health Faculty Publications by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact [scholarworks@gsu.edu](mailto:scholarworks@gsu.edu).

## NOTES2: Networks-of-Traces for Epidemic Spread Simulations\*

Sicong Liu and Yash Garg and K. Selçuk Candan

Arizona State University  
email: {sliu104, ygarg, candan}@asu.edu

Maria Luisa Sapino

University of Torino  
email: marialuisa.sapino@unito.it

Gerardo Chowell-Puente

Arizona State University  
email: gchowell@asu.edu

### Abstract

Decision making and intervention against infectious diseases require analysis of large volumes of data, including demographic data, contact networks, age-specific contact rates, mobility networks, and health-care and control intervention data and models. In this paper, we present our *Networks-Of-Traces for Epidemic Spread Simulations (NOTES2)* model and system which aim at assisting experts and helping them explore existing simulation trace data sets. NOTES2 supports analysis and indexing of simulation data sets as well as parameter and feature analysis, including identification of unknown dependencies across the input parameters and output variables spanning the different layers of the observation and simulation data.

### Introduction

Real-time and continuous analysis and decision making for infectious disease understanding and intervention involve multiple aspects, including (i) estimating transmissibility of an epidemic disease, such as influenza (Abubakar et al. 2012); (ii) forecasting the spatio-temporal spread of pandemic disease at different spatial scales (Merler et al. 2011); (iii) assessing the effect of travel controls during the early stage of the pandemic (Colizza et al. 2007); (iv) predicting the effect of implementing school closures (Wu et al. 2010); and (v) assessing the impact of pharmaceutical interventions on pandemic disease (Ferguson et al. 2005; Deodhar et al. 2014) through simulations. While highly modular and flexible epidemic spread simulation software, such as GLEaMviz (Van den Broeck et al. 2011) and STEM (STEM 2014), exist, these suffer from two major challenges that prevent real-time decision making:

- **Data and model complexity:** A sufficiently useful disease spreading simulation tool requires models, including demographic data, social contact networks, age-specific contact rates, local and global mobility patterns of individuals (Balcan et al. 2009; Merler and Ajelli 2014), epidemiological parameters for the infectious disease (e.g., infectious period), and control intervention data and models. Moreover, these dynamically evolve over time due to

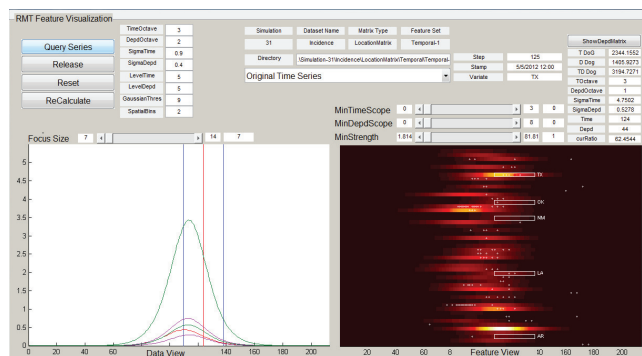


Figure 1: Simulation trace exploration interface of NOTES2

preventive actions taken by individuals and public health interventions, requiring continuous adaptation.

- **Complexity of the simulation and observation data:** Epidemic simulations track 10s or 100s of inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different resolutions. Moreover, generating an appropriate ensemble of stochastic epidemic realizations may require multiple simulations, each with different parameters settings corresponding to slightly different, but plausible, scenarios (Barrett, Eubank, and Smith 2005; Chao et al. 2010). Thus, running and interpreting simulation results (along with the real-world observations) to generate timely actionable results are difficult.

In this paper, we present *Networks-Of-Traces for Epidemic Spread Simulations (NOTES2)* to assist experts in exploring large simulation ensembles (Figure 1). The NOTES2 system supports

- *analysis and indexing of simulation data sets*, including extraction of salient multi-variate temporal features from the inter-dependent parameters, spanning multiple layers and spatial-temporal frames, driven by complex dynamic processes operating at different resolutions.
- *parameter and feature analysis*, including identification of unknown dependencies across the input parameters and output variables spanning the different layers of the observation and simulation data.

\*Supported by NSF grants #1318788 and #1339835.

## Related Works

**Temporal Data Analysis.** There are various multi-variate temporal data models, such as the *multi-variate structural time series* model (Harvey and Koopman 1997; Silva, Hyndman, and Snyder 2010).

Analysis of relationships (correlations, transfer functions, and causality) among time series is expensive (Reinsel 2003). A common representation of multi-variate data evolving over time is a tensor (multidimensional array) stream.

Tensors and tensor streams are often analyzed for their underlying structures through tensor decomposition algorithms (Carroll and Chang 1970; Harshman 1970; Tucker 1966). An alternative to tensor decomposition is to use probabilistic and generative models, such as Hidden Markov models (HMMs) and Dynamic Topic Modeling, DTM (Blei and Lafferty 2006).

A third alternative is to leverage AutoRegressive Integrated Moving-Average (ARIMA) and multi-variate ARIMA based analysis, which separates a time series into autoregressive, moving-average, and integrative components for modeling and forecasting (Mills 1990).

**Time Series Search.** In many applications, when comparing two sequences or time series, exact alignment is not required. Instead, whether two sequences are going to be treated as matching depends on how similar they are; thus, this difference needs to be quantified. This is commonly done through *distance* measures which quantify the minimum number (or cost) of symbol *insertions*, *deletions*, and *substitutions* needed to convert one sequence to the other. *Dynamic time warping (DTW)* distance (Ding et al. 2008; Keogh 2002; L.Ye and E.Keogh 2009; Yu et al. 2007), used commonly when comparing *continuous sequences* or time series can be thought of as a special case.

**Diffusion in Networks.** Kempe et al. were among the first teams who have investigated the problem of optimizing the network for maximum spread (Kempe, Kleinberg, and Tardos 2003). Watts and Dodds also studied the conditions under which nodes in a network become influential (Watts and Dodds 2007). (Chen, Wang, and Wang 2010) proposed a heuristic algorithm, based on local influence regions, to identify nodes in a network that maximize the spread of influence. In (Shakarian, Subrahmanian, and Sapino 2010), the authors focused on learning diffusion models and studying the impact of one node on the others in the network through reasoning with previously learned diffusion models, expressed via generalized annotated programs. (Leskovec et al. 2007) focuses on the related problem of optimal sensor placement to observe information cascades within the network, including disease outbreaks in a population, contaminant distribution within a water distribution network, and information flow within the blogosphere. (Kim, Candan, and Sapino 2012) noted that, while details differ, the various propagation models have two common properties: (a) *decay with distance*, and (b) *reinforcement*. Unfortunately, most models focus on the steady state of the propagation in the network and ignore the temporal dynamics of the diffusion itself. Moreover, most (if not all) of these works focus on a single parameter, whereas we need to track temporal

dynamics of multiple inter-dependent parameters.

## Networks-of-Traces for Epidemic Simulations

If effectively leveraged, models reflecting past outbreaks, existing simulation traces obtained from simulation runs, and real-time observations incoming during an outbreak can be collectively used for better understanding the epidemic's characteristics and the underlying diffusion processes, forming and revising models, and performing exploratory, if-then type of hypothetical analyses of epidemic scenarios.

There are five major types of data associated to epidemic spread simulations.

- *Network layers:* An epidemic simulation requires one or more layers of networks, from local and global mobility patterns to contact networks.
- *Disease models,* describing the epidemiological parameters relevant to a simulation and the parameter dependencies necessary in the computation of the disease spread.
- *Simulation traces:* For a given disease study, researchers and decision makers often perform multiple simulations, each corresponding to different sets of assumptions (disease parameters or models) or context (e.g. spatio-temporal context, outbreak conditions, interventions).
- *Disease observation traces:* These include real-world observations relating to particular epidemic, including the spread and severity of the disease and observations about other relevant parameters, such as the average length of recovery or percentage of infectious individuals that undergo pharmaceutical treatment.
- *External interventions:* In an outbreak, public health and disease control agencies implement various medical or social interventions, quarantines and/or school closures.

We collectively refer to these data (network layers, disease models, simulation traces, observation traces, and interventions) as the networks-of-traces (NT) data.

## Leveraging the NT Model for Disease Spread Simulation Understanding and Analysis

Epidemic spread simulations are complex. However, parameter dependencies and the network structures of the layers (e.g. mobility, social contact networks) are implicitly evident in the simulation traces and these carry temporal features (that may correspond to major changes in the underlying networks and/or temporal dynamics) that are robust against noise. The detection of these robust multivariate features constitutes the first step towards leveraging the NT data for understanding epidemics' characteristics and the diffusion processes, revising models, and performing exploratory, if-then type of hypothetical analyses of epidemic scenarios.

## Networks-of-Traces (NT) Feature Extraction

An NT data trace is multi-variate and the analysis of the relevant processes requires multi-variate temporal features spanning multiple inter-dependent trace parameters. Intuitively, a robust temporal feature in a multi-variate time series corresponds to a *multi-variate segment* of the series which significantly differs from its neighborhood. The multi-variate

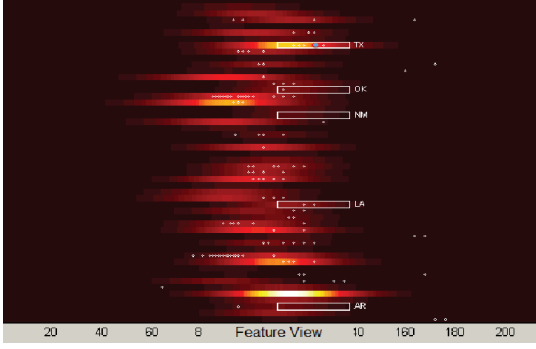


Figure 2: Each row corresponds to a time series of incidences for a sample epidemic simulation and each dot corresponds to the center of an identified multivariate feature

segment is represented by a center,  $\langle \mu_t, \mu_v \rangle$ , and a scope,  $\langle \sigma_t, \sigma_v \rangle$ . Intuitively,  $\mu_t$  marks the center of the segment in time and  $\sigma_t$  is the corresponding time interval. On the other hand,  $\mu_v$  is one or more nodes/variates of the graph on which the segment is centered and  $\sigma_v$  denotes all the graph vertices covered by the segment.

Figure 2 shows an epidemic simulation heatmap, where each row corresponds to a different state. In the figure, centers of identified features are highlighted by white dots. The figure also expands one of these robust features (tail end of the epidemic on a set of neighboring states): the center,  $\langle \mu_t = 125, \mu_v = \{TX\} \rangle$ , is marked with a blue dot and its scope,  $\langle \sigma_{time} = [111, 139], \sigma_v = \{AR, LA, NM, OK, TX\} \rangle$ , is visualized using rectangles.

**Robust Feature Detection** Let  $Y(t) = \langle Y_1(t), \dots, Y_m(t) \rangle$  be a multi-variate trace, from time  $t = 1$  to  $t = n$ . As in (Lowe 2004), we detect stable multi-variate features at the extrema of the scale space. However, unlike (Lowe 2004) (which operates on images with two ordered dimensions; i.e., rows and columns of pixels), extracting multi-variate features of the simulation trace  $Y$  (at various temporal and variate scales) requires detecting local maxima and computing gradients relative to not only the (ordered) time dimension, but also to the underlying variate graph.

**Time-and-Variate Smoothing.** We construct the scale space of  $Y(t)$  (corresponding to the versions of the series smoothed at different temporal and variate scales) relying on the following time-and-variate smoothing process:

- Let  $Y_i(t, s_t)$  indicate a version of uni-variate series,  $Y_i$ , smoothed with parameter  $s_t$ :  $Y_i(t, s_t) = G(t, s_t) * Y_i(t)$ , where  $*$  is convolution in  $t$  and  $G(t, s_t)$  is the Gaussian. Let  $Y(t, s_t) = \langle Y_1(t, s_t), \dots, Y_m(t, s_t) \rangle$  be a version of  $Y$ , where each uni-variate series is independently smoothed.
- Let us also define the variate smoothing function,  $S(\mathbf{R}, s_v, X) = [G(0, s_v)I + \sum_{(j=1)}^{\infty} 2 \times G(j, s_v)\mathbf{R}^j]X$ , where (a)  $\mathbf{R}$  is an  $m \times m$  matrix describing the variate dependencies, (b)  $X = \langle X_1, \dots, X_m \rangle$  is a  $m$ -vector, and (c)  $s_v$  is a variate smoothing parameter. Since  $G(j, s_v)$  approaches 0 quickly as  $j$  increases, the smoothing term in front of  $X$  can be approximated by a finite summation.

The time-and-variate smoothed version of  $Y(t, s)$  at scale  $s = \langle s_t, s_v \rangle$  is defined as  $\mathbf{Y}(t, s) = (\mathcal{H}_1(s); \dots; \mathcal{H}_t(s))$ , where  $\mathcal{H}_t(s) = S(H, s_v, Y(t, s_t))$  is the version of  $Y(t, s_t) = \langle Y_1(t, s_t), \dots, Y_m(t, s_t) \rangle$ , variate-smoothed at scale  $s_v$  at time instant,  $t$ .

**Iterative Scale Space Construction.** We construct the scale space by incrementally smoothing  $Y$  (both in time and variates) starting from an initial scale  $s_0 = \langle s_{t,0}, s_{v,0} \rangle$ . Let  $\mathbf{Y}_i(t, s)$  be a time-and-variate smoothed version of  $Y_i(t)$  at scale  $s = \langle s_t, s_v \rangle$ . Given a pair,  $k = \langle k_t, k_v \rangle$ , of time and variate scale multipliers, we add the three *scale-space neighbors* (or *ss-neighbors*) of  $\mathbf{Y}_i(t)$  into the scale space:

$$\begin{aligned} \mathbf{Y}_i(t, k \circ_t s) &\equiv_{def} \mathbf{Y}_i(t, \langle k_t \times s_t, s_v \rangle), \\ \mathbf{Y}_i(t, k \circ_v s) &\equiv_{def} \mathbf{Y}_i(t, \langle s_t, k_v \times s_v \rangle), \text{ and} \\ \mathbf{Y}_i(t, k \circ_{t,v} s) &\equiv_{def} \mathbf{Y}_i(t, \langle k_t \times s_t, k_v \times s_v \rangle). \end{aligned}$$

The process continues iteratively until maximum temporal and variate scales (bounded by the length of the simulation trace and the number of variates) are met.

**Local Extrema Detection.** For detecting extrema, for each  $\mathbf{Y}_i(t, s)$  in the constructed scale space, we compute

$$\begin{aligned} D_i^t(t, s) &= \text{abs}(\mathbf{Y}_i(t, s) - \mathbf{Y}_i(t, k \circ_t s)), \\ D_i^v(t, s) &= \text{abs}(\mathbf{Y}_i(t, s) - \mathbf{Y}_i(t, k \circ_v s)), \\ D_i^{t,v}(t, s) &= \text{abs}(\mathbf{Y}_i(t, s) - \mathbf{Y}_i(t, k \circ_{t,v} s)). \end{aligned}$$

Local extrema are identified by considering each  $\langle i, t, s \rangle$  triple and comparing  $\max(D_i^t(t, s), D_i^v(t, s), D_i^{t,v}(t, s))$  against the 78 *ss-neighbors* of  $\langle i, t, s \rangle$  in terms of time (*before, same time, after*), variates (*impacting, same variate, impacted\_by*), and scales (*smaller, same scale, larger*). Poorly defined extrema (i.e., an extremum that has a large principal curvature in one direction but a small one in the perpendicular direction) are eliminated.

**Feature Descriptor Creation.** Let us be given a triple  $\langle i, t, s \rangle$ . Let also  $N$  and  $M$  be two integers such that  $N \sim 3\sigma_t$  and  $M \sim 3\sigma_v$ . We create the local feature descriptor corresponding to this triple using a  $2N \times 2N$  matrix  $W$ : Let  $\mathbf{Y}_{(i,s)}$  be the time series  $Y_i$  at scale  $s$ ; then, for all  $-N < a \leq N$  and  $-N < b \leq N$ ,  $W[a, b]$  is defined as follows: (a) if  $b > 0$ ,  $W[a, b] = (\mathbf{R}^b \mathbf{Y}_{(i,s)})[t + a]$ ; (b) if  $b = 0$ ,  $W[a, b] = \mathbf{Y}_{(i,s)}(t + a)$ , and (c) if  $b < 0$ ,  $W[a, b] = (\mathbf{R}^{-1})^{|b|} \mathbf{Y}_{(i,s)}[t + a]$ .

Finally, we construct a  $(2u \times 2v \times c)$ -dimensional descriptor for the triple  $\langle i, t, s \rangle$  in the form of a gradient histogram based on the matrix,  $W$ : we sample  $c$  gradient magnitudes on the descriptor using a  $2u \times 2v$  grid superimposed on the matrix,  $W$ . A Gaussian weighting function is used to reduce the magnitude of elements further from the center.

## Feature Search and Alignment

Features extracted from a networks-of-traces data play important roles in the NOTES2 system. Here, we discuss how the similarity between two triples,  $\langle i_1, t_1, s_1 \rangle$  and  $\langle i_2, t_2, s_2 \rangle$ , and the corresponding descriptors,  $desc_1$  and  $desc_2$ , are computed in NOTES2. Depending on the use context, feature similarity has three major components:

Table 1: Target feature parameters

min target feature length	~ 5 time units
max target feature length	~ 40 time units
min target feature size	~ 2 hops
max target feature size	~ 10 hops
descriptor size	32 (= 2 × 2 × 8)

- *Descriptor alignment*: Since the feature descriptors are gradient histograms, their similarity is measured through a histogram similarity function (in the experiments, we use inverse of Euclidean distance).
- *Temporal alignment*: For temporal alignment between two features, we consider both the distance between the temporal centers of the features as well as the degree of overlap between the temporal scopes of the features.
- *Variate alignment*: For variate alignment, we consider both the distance between the variates in the underlying relationship graph as well as the degree of overlap between the variates within the scopes of the two features.

Depending on the application, we also consider alignments of (a) the *average amplitudes* and (b) *sizes* of the temporal and variate scopes of the two triplets. These various components of feature similarity are combined using a similarity merge function, such as *max*, *min*, or *product* based on the desired matching semantics.

## Evaluation

To assess whether the features extracted from epidemic simulations truly reflect the underlying *variate networks*, we created a set of simulations, using the STEM simulator, based on the US border network, where there is an edge between states if they share a border, and air network, which is a clique. For a given pair of transmission and recovery rates, we created 51 simulations (of length 213 units of time) assuming a different US state as the ground zero and recorded incidence rates<sup>1</sup>. We then extracted three sets of features from each simulation, using parameters in Table 1, and assuming different connectivity structures:

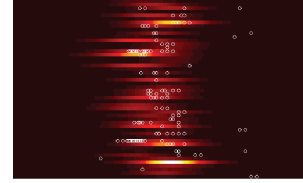
- *Border network (BN)*: For this case, we used the border network denoting states sharing borders.
- *Air network (AN)*: In this case, features are extracted assuming the air network (which is a clique).
- *Random network (RN)*: In this case, a random graph (with the same number of edges as the border network) is used for extracting features.

Given these features and their descriptors, we then computed the *confusion* for a simulation with ground zero state,  $gz_i$ , as  $confusion(gz_i) = AVG_{gz_j \neq gz_i} \frac{sim(gz_i, gz_j)}{sim(gz_i, gz_i)}$ . Here the similarity,  $sim(gz_i, gz_j)$ , between two simulations with ground zero states,  $gz_i$  and  $gz_j$ , is defined as  $\sum_{f \in features(gz_i)} sim(f, bestmatch_j(f))$ , where  $bestmatch_j(f)$  is the best matching feature to  $f$  in the simulation with ground zero state,  $gz_j$ . Temporal alignment parameters were set to be equal;  $\alpha_t = \beta_t = 0.5$ . Variate alignment parameters were set to  $\alpha_v = 0$  and  $\beta_v = 0.1$ , to avoid

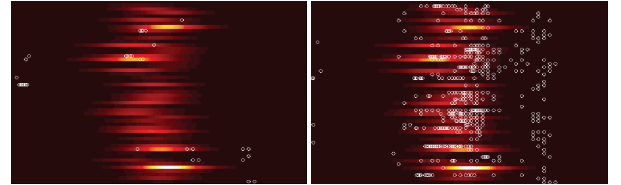
<sup>1</sup>Unless otherwise stated, we use the default STEM parameters

Table 2: Average confusions for simulations with different transmission and recovery rates.

<i>T.Rate</i>	<i>R.Rate</i>	<i>BN</i>	<i>AN</i>	<i>RN</i>
1.0	0.5	0.35	0.36 (1.14×)	0.44 (1.28×)
0.75	0.5	0.23	0.25 (1.06×)	0.32(1.37×)
0.25	0.5	0.37	0.63 (1.73×)	0.49 (1.33×)
1.0	0.25	0.44	0.45 (1.02×)	0.54 (1.23×)
0.75	0.25	0.39	0.45 (1.14×)	0.48 (1.23×)
0.5	0.25	0.29	0.34 (1.15×)	0.39 (1.32×)



(a) features extracted using the border network



(b) features using the air network (clique) (c) features using a random network

Figure 3: Centers of features extracted using different networks (source = “NJ”, trans. rate = 0.75 and rec. rate = 0.5)

penalizing the *wrong network* alternative. We use the product merge function to combine alignment scores.

Intuitively, large confusion implies poor differentiation power and, if the feature extraction process is effective, then we expect that (a) the overall confusion will be the lowest when using features extracted based on a network reflecting the underlying disease propagation, and (b) confusion will be the highest when we use an inappropriate network for feature extraction. Table 2 presents results for different transmission and recovery rates. As we see in this table, using the border network for feature extraction leads to least amount of confusion. Moreover, these results conform to our expectations listed above and Figure 3 helps see why: the (clique structured) air network ignores disease transmissions through land borders (especially when the transmission rate is too small for the flights to have a big impact on the epidemic’s diffusion) and, thus, misses useful features. Random networks, on the other hand, result in significant noise.

## Conclusions

In this paper, we presented our *networks-of-traces* model, which accounts for layers of disease networks (from local and global mobility patterns to contact networks), disease models, simulation and observation traces, and external interventions. The *Networks-of-Traces for Epidemic Spread Simulations (NOTES2)* system, based on this model, aims to assist experts in exploring large simulation trace data sets, through networks-of-traces feature analysis.

## References

- Abubakar et al., I. 2012. Global perspectives for prevention of infectious diseases associated with mass gatherings. *Lancet Infect Dis.* 12(1):66–74.
- Balcan et al., D. 2009. Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Medicine* 7(45).
- Barrett, C.; Eubank, S.; and Smith, J. 2005. If smallpox strikes portland. *Scientific American* 292(3).
- Blei, D. M., and Lafferty, J. D. 2006. Dynamic topic models. In *ICML'06*, 113–120.
- Carroll, J., and Chang, J.-J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika* 35.
- Chao, D.; Halloran, M.; Obenchain, V.; and Longini Jr., I. 2010. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput Biol* 6(1).
- Chen, W.; Wang, C.; and Wang, Y. 2010. TScalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD'10*.
- Colizza, V.; Barrat, A.; Barthélemy, M.; Valleron, A.; and Vespignani, A. 2007. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Comput Biol* 4(1).
- Deodhar et al., S. 2014. An interactive, web-based high performance modeling environment for computational epidemiology. *ACM TMIS* 5(2):7:1–7:27.
- Ding et al., H. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. In *VLDB08*, 1542–1552.
- Ferguson, N.; Cummings, D.; Cauchemez, S.; Fraser, C.; Riley, S.; Meeyai, A.; Iamsirithaworn, S.; and Burke, D. 2005. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature* 534(7046).
- Harshman, R. 1970. Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis. *UCLA Working Pap. in Phonetics* 16.
- Harvey, A. C., and Koopman, S. J. 1997. Multivariate structural time series models. *System Dynamics in Economic and Financial Models* 269–298.
- Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *KDD'03*, 137–146.
- Keogh, E. 2002. Exact indexing of dynamic time warping. In *VLDB02*, 406417.
- Kim, J. H.; Candan, K. S.; and Sapino, M. L. 2012. Impact neighborhood indexing (ini) in diffusion graphs. In *CIKM'12*, 2184–2188.
- Leskovec et al., J. 2007. Cost-effective outbreak detection in networks. In *KDD'07*.
- Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60(2).
- L. Ye, and E. Keogh. 2009. Time series shapelets: a new primitive for data mining. In *KDD09*.
- Merler, S., and Ajelli, M. 2014. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc Biol Sci.* 277(1681).
- Merler, S.; Ajelli, M.; Pugliese, A.; and Ferguson, N. 2011. Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: implications for real-time modelling. *PLoS Comput Biol.* 7(9).
- Mills, T. C. 1990. *Time Series Techniques for Economists*. Cambridge University Press.
- Reinsel, G. 2003. *Elements of Multivariate Time Series Analysis*. Berlin-Heidelberg: Springer-Verlag.
- Shakarian, P.; Subrahmanian, V.; and Sapino, M. L. 2010. Using generalized annotated programs to solve social network optimization. In *ICLP*, 182–191.
- Silva, A.; Hyndman, R. J.; and Snyder, R. D. 2010. The vector innovation structural time series framework: a simple approach to multivariate forecasting. *Statistical Modelling* 10(4):353–374.
- STEM. 2014. The spatiotemporal epidemiological modeler project. <http://www.eclipse.org/stem/>.
- Tucker, L. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31(3).
- Van den Broeck et al., W. 2011. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infect Dis.* 11(37).
- Watts, D., and Dodds, P. 2007. Influentials, networks, and public opinion formation. *J. of Consumer Research* 34(4).
- Wu et al., J. 2010. School closure and mitigation of pandemic (h1n1) 2009, hong kong. *Emerg Infect Dis* 16(3).
- Yu, D.; Yu, X.; Hu, Q.; Liu, J.; and Wu, A. 2007. Dynamic time warping constraint learning for large margin nearest neighbor classification. *Inf. Sci.* 18(13).