Spatial Data Mining to Support Pandemic Preparedness

Chris Bailey-Kellogg
Dept. of Computer Science
Dartmouth College
Hanover, NH 03755, USA
cbk@cs.dartmouth.edu

Naren Ramakrishnan
Dept. of Computer Science
Virginia Tech, Blacksburg
VA 24061, USA
naren@cs.vt.edu

Madhav V. Marathe
Virginia Bioinformatics Institute
Virginia Tech, Blacksburg
VA 24061, USA
mmarathe@vbi.vt.edu

ABSTRACT

Effective detection of and response to pandemic disease outbreaks require significant advances in data mining. Contributions to the recently held *SIAM DM 2006 Workshop on Spatial Data Mining* highlighted key challenges, directions, and progress in this context. We summarize here the main themes presented at the workshop as well as promising research directions for the data mining community.

1. INTRODUCTION

Pandemic diseases such as avian influenza cause extremely infectious disease outbreaks. Pandemic influenza viruses have demonstrated their ability to spread worldwide within months or even weeks, and to cause infections in all age groups. While the ultimate number of infections, illnesses, and deaths is unpredictable, and could vary tremendously depending on multiple factors, it is nonetheless certain that without adequate planning and preparation [13], an influenza pandemic has the potential to overwhelm current public health and medical care capacities at all levels. Controlling the spread of a pandemic requires early detection via appropriate surveillance [1], along with implementation of appropriate responses [6] (e.g., isolation of cases, quarantine of contacts, antiviral drug treatment and prophylaxis). These needs directly motivate research in spatial data mining for a time-varying network capturing collocation and effective contact patterns [5].

The SIAM DM 2006 Workshop on Spatial Data Mining (http://www.cs.dartmouth.edu/~cbk/sdm06/) provided a forum for explorations into these challenges. To focus the discussion, a synthetic dataset of disease evolution in the city of Portland, Oregon was provided by the Virginia Tech Network Dynamics and Simulation Science Laboratory [9]. The five regular papers accepted for presentation at the workshop [3; 7; 11; 12; 14] showcased a variety of data mining studies performed on this dataset, e.g., model-based data aggregation, mining spatial interaction patterns, predicting infection risks, designing containment policies, and process-driven spatial and network aggregation. Two additional short papers [8; 15] addressed spatial data mining challenges more generally. The goals of this report are two-fold: to summarize the main themes resulting from the workshop and to bring to the attention of the larger data mining community the challenging problems arising in the context of pandemic preparedness.
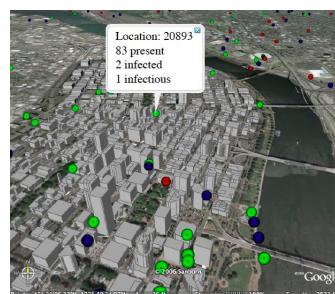

Figure 1: Visualization of synthetic dataset.

2. SYNTHETIC DATA FOR PANDEMIC PREPAREDNESS STUDIES

Effective pandemic disease modeling must necessarily take into account multiple aspects of geography, epidemiology, social structures, and network dynamics; however, historical datasets that provide adequate coverage over all relevant aspects are very scarce. There is hence an established practice of using synthetic datasets and mathematical models to understand the course of disease outbreaks and to design effective policies. Furthermore, synthetic datasets protect the privacy of individuals and other proprietary and sensitive information. While real data are likely to be more limited than the omniscient view afforded by synthetic data, the synthetic data can serve as an oracle in studies of observability, and in development of appropriate detection and active sampling techniques. In particular, we can assess relationships between factors that we would like to observe and factors that can actually be observed. Similarly, studies employing synthetic data can identify what information (type, quantity, etc.) is required for effective response policies.

The supplied dataset tracks a set of synthetic individuals in Portland (Fig. 1) and, for each of them, provides a small number of demographic attributes (age, income, work status, household structure) and daily activities representing a normative day (including places visited and times). The city itself is modeled as a set of aggregated activity locations, two per roadway link. A collection of interoperable simulations—modeling urban infrastructure, people activities, route plans, traffic, and population dynamics—mimic the time-dependent interactions of every individual in a regional area. Disease dynamics are captured by a coupled probabilistic timed transition system whereby the state of health of a person can change depending on the health and duration of his or her contacts in the underlying social network. This form of 'individual modeling' is in contrast to

the traditional approach of estimating gross reproductive numbers by uniform mixing models over the entire population. In particular, it provides a bottom-up approach mirroring the contact structure of individuals and is naturally suited for formulating and studying the effect of intervention policies. For more details, please see the EpiSims project (`http://ndssl.vbi.vt.edu/episims.html`; [4]), part of the TRANSIMS simulation infrastructure (`http://ndssl.vbi.vt.edu/transims.html`; [2]).

3. CHALLENGES

Several levels of analysis must come together to successfully address the data mining challenges in pandemic preparedness. First, from the synthetic dataset, we must develop and model a time-varying spatial-social network capturing collocation and effective contact patterns. Second, we must conduct model-based data aggregation to identify the onset of disease and other qualitative indicators of disease spread. Third, we must identify critical individuals and critical locations, in order to support targeted vaccination and targeted detection goals (respectively). We discuss some of the work discussed at the workshop for these various challenges.

3.1 Modeling Spatial Data

One of the key themes brought out by the workshop is a multi-faceted definition of what it means to be 'spatial', and how to appropriately model the data under that notion of spatiality. Modeling disease outbreaks necessitates capturing not only the geographic context but also the induced neighborhoods caused by people's movement patterns and visitation patterns as a function of time. There is a complex interplay between 'spatial' in the geographic sense and 'spatial' according to distances in a social network—propagation in one context appears as discontinuous 'jumps' in the other.

A wealth of literature exists on how diseases propagate through spatial and social channels, both at the epidemiological level and at the network modeling level. At the same time, data mining techniques can integrate multiple views of locality and spatiality. For instance, Gaussian process (GP) methods that work with a prior on covariance structures [10] can be fruitfully interleaved with techniques that define covariances using an underlying graph model. Savell and Chung [11] propose an approach using the related Gaussian Random Fields (GRFs) to model stochastic diffusion of disease state on the underlying network. This supports disease state prediction for unlabeled nodes, and thereby could account for real-world limitations in available information.

Several other generalizations of 'spatial' were utilized in the presented work. The original study by Marathe and co-workers [5] included graph models such as people-people interactions (by way of collocation) and location-location interactions (by way of shared visitors). Guo [7] studies the relationship between such social localization and the underlying spatial localization. Tatikonda et al. [12] use these graph models as the basis for developing containment policies, discussed further below. Chen et al. [3] adopt a similar approach in analyzing Portland's electrical network, which contains a mixture of short and long distance interconnections, as well as notions of proximity to sources (the generation system) and sinks (the consumers). Zarnani and Rahgozar [15] start with a similar neighborhood graph representation, while Jin et al. [8] integrate metric spatial and temporal information. Vucetic and Sun [14] further account for location type (school, work, home, etc.) as a key attribute defining the spatial context.

3.2 Identifying Meaningful Spatial Structures

At the heart of spatial data mining is the uncovering of multi-level structures that enable new insights into the underlying data (and here, strategies for responding). This is possible because spatial data exhibit similarities and continuities at multiple levels. Multi-level approaches also allow analyses to scale to large datasets, as used in studies such as this one. Guo [7] seeks to detect interaction patterns in a multi-level approach, using a combination of sampling (while preserving overall interaction structure), clustering (bringing in geographic information), and projection (focusing on strong connections). Savell and Chung [11] uncover multi-level representations by identifying phase transitions. Thus temporal information (disease evolution and propagation) provides insights into the spatial structures.

Other types of application-dependent spatial structures are also highly illuminating. For example, consider the detection of vulnerabilities and criticalities. A person or location might be considered highly vulnerable based on dense connections, and might be considered highly critical based on anticipated 'downstream' effects. Guo [7] suggests using bridges between clusters in the interaction graph as critical points for early detection of pandemics. Criticalities are also obviously useful in pandemic response strategies. Structures that directly feed into response policies include profiles capturing similarities in local graph structure, used by Tatikonda et al. [12]; aggregation based on location type, used by Vucetic and Sun [14]; and temporal synchrony in highly-connected sites, used by Savell and Chung [11]. Zarnani and Rahgozar [15] look for trends in spatial data, using ant colony algorithms in postulating and evaluating paths through spatial neighborhood graphs. In the domain of traffic monitoring, Jin et al. [8] seek to identify anomalies (traffic incidents) by learning and monitoring spatio-temporal profiles of traffic flow.

3.3 Developing Control Policies

Ultimately the goal of mining the synthetic dataset is to design actionable policies for preventing and containing diseases, and many papers focused on the key sensing and planning issues. Issues of controllability and observability, with respect to specific resource constraints (cost, physical feasibility, robustness, etc.), are crucial here.

The paper by Tatikonda et al. [12] investigates various containment policies for transmissible diseases, including random, contacts-, sociability-, profile- and location-driven vaccination. These policies are variously based on either the people-location activities graph or the collocation network between people. Interestingly (and beneficially), they found that response can be delayed for quite some time, due to detection lag, with relatively small impact on total infection. The phase transitions found by Savell and Chung [11], on the other hand, warn that rapid order-of-magnitude shifts happen after the initial build-up phase.

Chen et al. [3] approach control design with a top-down strategy, using disaggregation of overall indicators down to individual locations, and determining if the disaggregation respects spatial proximity constraints. Although they focus on modeling electricity demand in the synthetic dataset, it is easy to see how this approach can be extended to disaggre-

gating other global indicators of disease outbreaks. Savell and Chung [11] likewise propose using global information, via their GRF model, in an active sampling strategy. Assessment of risk guides the active sampling, either towards nodes useful for observation (where classification risk is highest) or control (where infection risk is highest). Vucetic and Sun [14] also take a risk-based approach, using location aggregation (type of activity) to provide input for classifiers that predict infection risk. The first few generations of disease spread provide the training data for predicting the future course of the pandemic (and thereby appropriate responses).

4. CONCLUSION

Data mining in support for pandemic preparedness is an important and rich application area, with many significant research challenges. Summarizing and adding to the issues identified above, we identify six key areas where data mining research must make progress:

Algorithms for fast computations of multi-level network properties induced by spatial-social data; in particular, provable approximations for estimating expansion, between-ness, and community structure, and tracking such properties across time-indexed snapshots.

Integrating model-driven methods with spatial mining, e.g., combining a model for disease spread with a method for detecting critical individuals and locations; or using data mining to derive a social distancing policy or to formulate quarantine procedures.

Disaggregation on demand, i.e., determining a small set of multi-level aggregates (among the multitude of possibilities) to be stored as sufficient statistics, thus allowing other microscopic parameters to be re-generated on demand.

Co-evolving epidemic policy, simulation, and mining; unlike passive observation of data to derive targeted sampling policies that model a static dataset, implementing an intervention policy can fundamentally change the course of future simulation runs, thus making data mining an integral part of the simulation-based model.

New objective functions for active data mining, that mimic targeted detection and targeted vaccination goals in epidemiological modeling, and, in this manner, close the monitor-simulate-mine loop.

Support the view of simulation models as procedural representations of large datasets; thus allowing the rich modeling literature to be harnessed for data mining goals.

We hope these proceedings of the SIAM DM 2006 Workshop on Spatial Data Mining serve as an impetus toward establishing a new thrust in the practical problem of pandemic modeling and result in a consolidation of ideas as well as a renewed bearing in spatial data mining research.

Acknowledgments: We sincerely thank the group members of NDSSL for their help in creating the synthetic datasets.

5. REFERENCES

[1] C. Barrett, J. Smith, and S. Eubank. Modern Epidemiological Modeling. *Scientific American*, Feb 2005.

[2] C.L. Barrett, R.J. Beckman, K.P. Berkbigler, K.R. Bisset, B.W. Bush, S. Eubank, K.M. Henson, J.M. Hurford, D.A. Kubicek, M.V. Marathe, P.R. Romero, J.P. Smith, L.L. Smith, P.L. Speckman, P.E. Stretz, G.L. Thayer, E. Van Eeckhout, and M.D. Williams. TRAN-SIMS: Volumes I, II, III, and IV. Los Alamos Unclassified Reports 00–1724, 00–1725, 00–1766, and 00–1767, Aug 2004.

[3] J. Chen, V.S. Anil Kumar, A. Marathe, and K. Atkins. Model Based Spatial Data Mining for Power Markets. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 6 pages; Bethesda, MD.

[4] S. Eubank, V.S. Anil Kumar, M. Marathe, A. Srinivasan, and N. Wang. Structure of Social Contact Networks and their Impact on Epidemics. AMS-DIMACS Special Volume on Epidemiology, 2006. to appear.

[5] 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*, Vol. 429:180–184, May 2004.

[6] N.M. Ferguson, D.A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D.S. Burke. Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia. *Nature*, Vol. 437:209–214, Sep 2005.

[7] D. Guo. Mining and Visualizing Spatial Interaction Patterns for Pandemic Response. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 6 pages; Bethesda, MD.

[8] Y. Jin, J. Dai, and C.-T. Lu. Spatial-Temporal Data Mining in Traffic Incident Detection. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 5 pages; Bethesda, MD.

[9] NDSSL. Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 1.0. Technical Report NDSSL-TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute (VBI), Virginia Tech, 2006. Available at http://ndssl.vbi.vt.edu/opendata/.

[10] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[11] R. Savell and W. Chung. Process Driven Spatial and Network Aggregation for Pandemic Response. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 7 pages; Bethesda, MD.

[12] S. Tatikonda, S. Mehta, and S. Parthasarathy. Containment Policies for Transmissible Diseases. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 7 pages; Bethesda, MD.

[13] US Department of Health and Human Services. HHS Pandemic Influenza Plan. http://www.hhs.gov/pandemicflu/plan/, Nov 2005.

[14] S. Vucetic and H. Sun. Aggregation of Location Attributes for Prediction of Infection Risk. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 5 pages; Bethesda, MD.

[15] A. Zarnani and M. Rahgozar. Mining Spatial Trends by a Colony of Cooperative Ant Agents. In *Proc. SIAM DM 2006 Workshop on Spatial Data Mining*, Apr 2006. 5 pages; Bethesda, MD.