Process Driven Spatial and Network Aggregation for Pandemic Response

Robert Savell and Wayne Chung
Dartmouth College, Thayer School of Engineering

02/27/2006

Abstract

Phase transitions in measures of cluster connectedness may be used to identify critical points in the propagation of an epidemic. These critical points reflect order of magnitude shifts in network properties and thus define appropriate regions for aggregation in the evolving socio-temporal portrait. Analysis of pre- and post- transitional images at these critical points can define principle axes of propagation and establish the appropriate local scale (aggregate level) for resource allocation strategies. Semi-supervised learning techniques based on Gaussian random fields allow prediction of infectious spread to unlabeled entities. Projections of disease propagation inform allocation strategies for intelligent targeting of response resources to the most vulnerable locations in the unlabeled network. Strategies for perimeter sensing and establishment of containment boundaries may also be defined in terms of classification risk of the unlabeled nodes.

1 Introduction:

Channels of epidemic propagation possess a natural multi-scalar structure, making the area of pandemic response and control an obvious target for hierarchical aggregation techniques. Theoretical challenges arise in the application of traditional techniques for spatial aggregation due to the interplay of spatio-temporal factors with social processes more naturally defined on the social network. This complexity results in a rich problem domain supporting the examination of practical and theoretical extensions of the spatial aggregation framework. The potential social benefit of the application provides additional incentive for the development of theoretically sound and practically effective solutions.

Several of the more challenging issues for the development and application of aggregation techniques in the context of pandemic response are:

- The complex interplay of spatial and social network structure in definition of propagation boundaries.

- The need to develop an evolutionary portrait of the pandemic, rather than a traditional stationary snapshot.

- The difficulty of defining global parameterizations of scale in an organically evolved spatio-social substrate such as a city environment.

- The need for information directed sampling and resource allocation strategies, due to the difficulty of data acquisition and implementation of response strategies in a real environment.

The appeal of multi-scalar techniques such as spatial aggregation lies in the recursive reduction of informational and computational complexity. In the aggregation process, the output of local processes defined on a lower scalar level are aggregated and passed on to the next higher level where an analysis of similar complexity is performed. The inter-scalar definition of the aggregate hierarchy is designed to capture order of magnitude shifts in behavior and structure. However, in the case of an epidemic propagating through a social network, it may be difficult to establish appropriate scalar levels for analysis of the spatial interactions of the processes. In fact, the concept of spatial neighborhood itself must be reconsidered, given the natural tendency of social processes to jump from region to region with shifts in behavioral context (for example, when an entity travels from home to work to school).

Rather than attempt to wrest a useful spatial decomposition of epidemic transmission channels from the tangle of temporally and contextually interleaved contact events which define the potential transmission pathways in a dynamic city environment, we shall adopt, instead, a process driven methodology. In this framework, the temporal behavior of the propagating epidemic, rather than merely adding layers of extra complexity to the analysis, may be viewed as providing temporal cues for segmentation of the evolutionary profile of the epidemic. By pinpointing order of magnitude shifts in the neighborhood structure of the infected network, a spatio-temporal or socio-temporal portrait of the structure of the propagating epidemic may be developed, with the dominant mechanism and associated scalar region defined for each phase transition. Given the pre- and post- images of a socio-temporal locality associ-

ated with a phase transition, the principle spatial vectors supporting the local infectious transmission may be identified. In the case of a simulation or a post-mortem analysis of an outbreak, in which complete information is assumed, a direct analysis may be performed on the fully labeled data set. Alternatively, in a realistic setting with limited infection data and sparse information as to social network interactions, exploratory and preventive resources may be directed via a combination of semi-supervised learning and active sampling techniques based on a gaussian random field description of classification potentials in the labeled and unlabeled space.

2 Background:

As an example of a spatial strategy for pandemic response, Fergusen et al. [1] demonstrates that a rapid response strategy concentrated in elliptical spatial regions of appropriate size about identified instances of influenza effectively halts the spread of the disease in a simulated outbreak in Southeast Asia. However, the simulated region is primarily rural and population density is treated as relatively sparse and homogenous. The current task is to examine a simulated outbreak in the more complex environment of the city of Portland. At this level of detail, the characteristics of the social network structure are much more critical to determining the course of infection, and the complexities of typical spatio-temporal interactions of spatial entities tends to make definition and isolation of localities difficult.

Since there is minimal cost associated with connection maintenance in a network of individuals infected in a pandemic scenario, the theory of "small-world" networks [5, 6]suggests that nonlinear phase transitions or order of magnitude shifts in certain measures of network connectivity or path length may be used to identify critical points in the temporal evolution of the pandemic. Analysis of the contact events and implied network connectivity provides a basis for defining the local attributes and spatial extent of the infectious expansion. Appropriate measures for monitoring, quarantine, population dispersal may then be applied at the proper scale.

When attempting to identify order of magnitude shifts in the structure of the infected network, simple, locally defined heuristic measures should be sufficient. Examples of two possible measures for use in cluster detection— characteristic path length and connection density— are described in [11, 12]. M.E. J. Newman [10] develops a useful measure for fast detection of community structure, which we describe in detail in Section 3.2.

Since it is unlikely that a complete description of local connectivity will be available in a practical

setting, predictions of unseen links and potential link formation may be informed by principles of preferential attachment and nodal influence as suggested by the (independent) work of Newman and Kleinberg [7, 8, 9]. Similar assumptions of increased likelihood of infection with proximity of individuals in the social network substrate suggest a diffusive or random walk model to describe the local mechanism for disease propagation.

Given this preferential model for social network formation and disease transfer, a method for stochastic labeling of unclassified nodes in the social network may be defined. Diffusive propagation of classification labels in a sparsely labeled undirected graph is addressed by Zhu, Lafferty, et al. [3, 4]. The semi-supervised learning and active sampling techniques based on assumptions of a gaussian random field defined on the nodes of the social network provide the foundation for development of practical method for allocating resources, given sparse knowledge of the social network and sparse reporting of infectious events. The methods in [3, 4] are closely related to techniques based on gaussian processes based active sampling for spatial aggregation methods, as described in [2]. Zhu et al. have the advantage in the current context; however, that the random field is defined in terms of the nxn connectivity matrix of the social network nodes rather than on a spatial grid.

3 Methodology

One of the most interesting aspects of the current problem is the fact that the social channels supporting pandemic propagation may be viewed as an evolving substrate. Where traditional spatial aggregation assumes pattern analysis on a fixed lattice, an aggregation technique operating on a social network must cope, in the absence of perfect information about the spatio-temporal social network structure, with a continuously evolving infrastructure of transmission. Our proposed methodology seeks to turn this difficulty into an advantage by using the nonlinear phase transitions associated with the connective properties of the densifying network of infected entities to identify critical socio-temporal regions in the dendogram.

As suggested by dynamic social network models positing preferential attachment [5, 6, 7, 8, 9] , we expect the probability of infection for a particular node to be a function (often linear) of the amount of contact the node has with infected neighbors. The dynamics of network organization driven by mechanisms of preferential attachment tend to exhibit nonlinearities or phase transitions in several connectivity measures as networks densify and network connectivity transitions from sparse to dense [5, 6]. Phase transitions are predicted, even in instances where the maximum neighborhood size per node

is restricted [7].

By tracking local measures of network connectivity we may readily identify critical moments in the organization of the infected network. These critical points correspond to order of magnitude shifts in structural density of the network, and thus reflect the optimal points for qualitiative aggregation. Analysis of the pre- and post-transition network images yields the spatial components associated with the expanded infectious cluster.

In this methodology, clustering of the dendogram is defined as a temporally evolving hierarchal agglomeration technique. And, in an analogy drawn from the path planning literature, the edges connecting well defined clusters in the dendogram are exploratory, while the connections formed during the nonlinear densification or consolidation phases may be considered exploitive.

3.1 Local measures for clustering.

Several possible measures for establishing cluster density or relative connectedness are possible. One common measure, characteristic path length (CPL) [11, 12] measures the average shortest path link distance between pairs of nodes. An attractive property of the measure is its relative insensitivity to the number of nodes in the graph; however, the metric may be too computationally intensive for establishing large clusters. Another common measure is the connection density, defined as the proportion of the total number of edges in the graph vs. the total possible edges. This measure is much easier to calculate than CPL, but is too sensitive to the number of nodes to be reliable indicator of connectedness.

A preferable measure of community structure is due to M.E.J. Newman [10] and compares the number of intracluster edges with the number of edges which would fall in the cluster if all edges incident to the cluster were placed randomly. With e_{ii} defined as the fraction of edges in the network whose endpoints both correspond to nodes in cluster i and e_{ij} defined as $1/2$ of the fraction of edges with exactly one endpoint in cluster i, the local density estimate for cluster i is given by:

$$Q_i = e_{ii} - e_{ij}^2$$

Summing over all clusters yields a global measure of community structure. The measure has several attractive properties including ease of computation. The measure qualitatively captures the essence of the clustering property. In addition, it is defined in terms of the actual edges incident upon a cluster, rather than potential edges— making it less sensitive to variations in cluster scale than the density measure. By thresholding this clustering measure, we may define a simple heuristic for iterative clustering as nodes are added to the evolving infectious network. The heuristic need only support

three monotonic behaviors in the clustering process.

- Addition of new clusters— exploratory edges connect sparse regions to the infected network.

- Intra-cluster densification.

- Inter-cluster densification.

Critical points in the propagation of the epidemic are identified during the densification process by spontaneous aggregation of clusters or by a rapid change in the local connectivity measure.

3.2 Spatial Aggregation, Link Prediction and Active Sampling at Critical Points.

Identification of critical points in the evolution of the infected network provides a roadmap in the socio-temporal space for application of spatial aggregation strategies. The nonlinear rate of coalescence at critical points effectively discretizes the evolutionary process. In the presence of complete information, as in a simulated enviroment, statistical forensics on the pre- and post- images of an infected cluster may be applied to determine the dominant spatio-temporal channels which were responsible for the local densification. Assuming complete information, this straightforward forensic analysis can yield a qualitative portrait of the expected evolution of transmission channels which is useful for grounding the development of practical resource allocation strategies.

While a forensic analysis can provide a useful framework for long term planning, a practical real-time response strategy must also consider issues of prediction. Obviously, a response strategy should not wait for physical phase transitions before determining scope of response, since these transitions correspond to order of magnitude escalations in infection. An effective response strategy should strive both to predict the location and scope of potential escalations and to project an appropriate firewall for containing the disease. Thus, effective prediction requires a probabilistic assessment of the potential risk of infection for nodes at the frontier of the infected network.

When considering projections onto unlabelled regions, many practical issues arise. In contrast to our simulated environment, in an actual pandemic situation, information on infected instances will likely be quite sparse, as will information as to the social interactions of affected individuals. A useful strategy for practical allocation of response resources must include a mechanism for defining expected disease classification labels on unsampled nodes, as well as a strategy for actively sampling infectious neighborhoods in the social network. Active sampling strategies serve several functions in this context including— identification of loca-

tions at immediate risk of infection and allocation of resources toward mapping the social network in order to establish the probabilistically defined frontier of the propagating infection.

Recent work attempts to exploit un-labeled sample points for spatial aggregation [2] by establishing a stochastic assignment of unlabeled regions via a gaussian process. As previously noted, the tendency of social network interactions to defy spatially defined neighborhood relationships suggests that stochastic methods should be designed to operate on the social interaction matrix. The work of Zhu et. al. [3, 4] provides a natural framework for stochastic classification and active sampling in the current context.

Given a graph defining connectivity of a collection of sparsely labeled nodes, a semi-supervised (incorporating labeled and unlabeled data) learning problem may be formulated in terms of a Gaussian random field on the graph. The minimum energy configuration of the gaussian random field, given the constraints defined by the node labelings and the connectivity matrix corresponds roughly to the equilibrium state of a diffusion process propagating in the unlabeled regions of the matrix. This interpretation is particularly suited to the current context, with the label diffusion reflecting the propagation of infected particles via a random walk process in the unexplored regions of the social network.

A brief description of the semi-supervised learning process described in [3] elucidates the utility of the gaussian kernel method. Placing the nodes x of the cluster of interest in $\mathbf{R}^m$ we define a weight matrix:

$$w_{ij} = \exp\left(-\sum_{d=1}^{m} \frac{(x_{id} - x_{jd})^2}{\sigma_d^2}\right)$$

where x_{id} is the d-th component of instance x_i represented as a vector $x_i \in \mathbf{R}^m$. (As a practical consideration, we expect the spatial representation of the data to be further differentiated according to modality of the spatial interactions— that is, home, work, school, etc.).

In the learning step, the σ_d's are fit using both labeled and unlabeled data via gradient descent in the hyperparameter space. Since we are dealing with unlabeled data, the usual optimization criterion— maximization of likelihood of the labeled data— does not apply. Instead, the *average label entropy* is used as a heuristic criterion. This is reasonable, since the space of low entropy labelings achievable by harmonic minimization is relatively small. As a practical side effect, the process of fitting these hyperparameters serves as a feature selection mechanism, establishing the orientation and magnitude of the principle vectors of infectious transmission.

The active sampling method in [4] selects unlabeled

points for classification which will lead to maximum reduction in classification risk (the estimated generalization error of the classifier). In the pandemic application, there are several potential goals for active sampling. A strategy might potentially target some resources toward unlabeled nodes with the maximum risk of infection (minimum risk of classification = 0) along principle vectors of transmission. Alternatively, epidemic firewalls may be defined by placing resources at maximal classification risk boundaries, in order to monitor the spread of the disease to new areas and to pre-empt order of magnitude jumps in the scale of infected regions.

4 Synopsis:

In conclusion, the proposed methodology for applying spatial aggregation techniques to the context of pandemic response and control may be summarized as follows:

1. Monitor local measures of network connectivity to identify phase transitions signaling critical points (order of magnitude scalar shifts) in propagation of the disease.

2. Maintain an evolving dendogram of the infected social network with cluster points defined socio-temporally at critical points identified in Step 1.

3. For each active cluster in the dendogram, establish critical channels and scale of propagation by comparison of pre- and post- transitional images. These channels may be determined via direct evidence in a forensic analysis or, in a practical response strategy, via a semi-supervised learning technique incorporating unlabeled data.

4. Employ active sampling in the semi-supervised framework to monitor cluster boundaries and to apply resources to high threat nodes along principle vectors of transmission.

References

[1] N. 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* **437**, pp. 209–214, 2005.

[2] N. Ramakrishnan, C. Bailey-Kellogg, S. Tadepalli, and V. Pandey, "Gaussian processes for active data mining of spatial aggregates.," in *SDM*, 2005.

[3] J. D. L. X. Zhu, Z. Ghahramani, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. 20th International Conference on Machine*

Learning (ICML'03), pp. 912–919, AAAI Press, January 2003.

[4] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *Proc. of the ICML Workshop on the Continuum from Labeled to Unlabeled Data*, pp. 58–65, AAAI Press, 2003.

[5] D. J. Watts, *Small Worlds*, Princeton University Press, Princeton, NJ, 1999.

[6] R. Albert and A. L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics* **74**, pp. 47–97, 2002.

[7] E. M. Jin, M. Girvan, and M. E. J. Newman, "The structure of growing social networks," *Physical Review Letters E* **64**, 2001.

[8] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pp. 556–559, 2003.

[9] D. Kempe, J. M. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks.," in *ICALP*, pp. 1127–1138, Springer Verlag, 2005.

[10] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E* **69**, p. 066133, 2004.

[11] T. R. Coffman and S. E. Marcus, "Pattern classification in social network analysis: A case study.," in *Proceedings of the 2004 IEEE Aerospace Conference (Big Sky, MT, Mar. 2004).*, 2004.

[12] T. R. Coffman and S. E. Marcus, "Dynamic classification of groups using social network analysis and hmms," in *Proceedings of the 2004 IEEE Aerospace Conference (Big Sky, MT, Mar. 2004).*, 2004.