Trade-offs Between Agility and Reliability of Predictions in Dynamic Social Networks Used to Model Risk of Microbial Contamination of Food

Artur Dubrawski, Purnamrita Sarkar and Lujie Chen
Auton Lab, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
awd@cs.cmu.edu, psarkar@cs.cmu.edu, karenchen@cmu.edu

Abstract—This paper evaluates trade-offs between agility and reliability of predictions arising due to sparseness of data modeled with dynamic social networks. We use real field data from food safety domain to illustrate the discussion. We model food production facilities as one type of entities in a social network evolving in time. Another type of entities denotes various specific strains of Salmonella. Two entities are linked in the graph if a microbial test of food sample conducted at the specific food facility over specific period of time turns out positive for the particular pathogen. We use a computationally efficient latent space model to predict future occurrences of pathogens in individual facilities. Empirical results indicate predictive utility of the proposed representation. However, sparseness of data limits the attainable agility of predictions. We identify exploiting recency of data and using the known patterns in it, such as seasonality, as plausible means of battling the challenge of sparseness.

Keywords—dynamic social networks; latent space models; food safety.

I. INTRODUCTION

Food-borne diseases pose a significant burden on societies and economies worldwide. It is estimated that in the United States each year there are approximately 76 million cases of such illnesses which require 325,000 hospitalizations and result in 5,000 deaths. A significant contribution to those gruesome statistics is caused by Salmonella, a bacteria causing acute gastrointestinal symptoms in 40,000 Americans annually, while 1% of cases of are terminal [1].

Some of those infections are attributable to preventable contamination of food at production facilities. Federal and local food safety and public health offices undertake widespread efforts to mitigate risks involving production, transportation, preparation and consumption of food to minimize exposure of the public. The spectrum of efforts includes routine microbial testing of samples of food taken at subsequent processing steps. The collected data can then be analyzed to e.g. survey the processes of transmission of pathogens, to assess efficacy of countermeasures, or to optimally allocate investigative resources to proactively respond to threats.

In this paper we focus on the analysis of microbial testing data collected at food factories. Such data can be used in the context of statistical process control to detect conditions indicating departures from sanitary safety bounds [2]. It can also be used to support trace-back investigations of identified clusters of cases of disease [3]. Predictive analytics can be used to estimate risk of microbial contamination of food at processing establishments based on their record of recent sanitary inspections [4]. Those approaches usually assume independence of data collected at individual food factories and they aggregate microbial test data to the level of individual bacteria. Recently, we tested utility of social network models to predict risk of microbial contamination of food at specific establishments based on their historical performance and on knowledge of past and current performance of their peers [5]. We have also used a dynamic social network paradigm to model such data [6]. Empirical evaluation of social network analysis applied to predicting results of microbial testing of food shows utility of relaxing the assumption of independence of food establishments. The results also indicate appropriateness of the graph-based representation for handling sparseness of microbial testing data.

In this paper we discuss the impact of the width of temporal window of analysis on accuracy and agility of predictions. On one hand, the more frequent prediction cycles, the less latent (more agile) the monitoring process can be. On the other hand, the narrower the window width, the lower the expected count of occurrences of specific microbial findings per food factory, hence the greater sparseness of data per each training time step and, consequently, the harder it is to obtain reliable predictions.

The following section of the paper introduces the dynamic social network model we use and explains its application to modeling of the microbial testing data. Then we introduce our experimental setup and discuss results of empirical evaluation of the proposed approach.

II. DATA REPRESENTATION

The key hypothesis behind our research poses that similar food factories tend to have similar histories of food safety problems. If it holds, we can use it to create predictive models for groups of similar factories. Such models would then likely yield greater accuracies than the alternative which treats all establishments as a single group. They should also
outperform, in terms of statistical reliability, the models that could be developed separately for individual factories. That is because the size of the sample of training data for a group would exceed all of the sizes of data collected for the individual entities.

There are several types of relationships between food factories which may be used as grouping factors. Fundamental criteria include shared structures of supply chain, corporate membership, production profiles, sizes, geographic locations, product distribution patterns, etc. Such information is potentially very useful, but it is not always available to food safety or public health agencies. Therefore in our research we primarily rely on phenomenological factors such as observed co-occurrences of specific serotype Salmonella positives at two or more food establishments during the same period of time. Such information can be readily extracted from the routinely collected data. It can be considered a weaker evidence of true relationships between establishments than what could be inferred using the aforementioned fundamental features. However, it is still reasonable to assume that truly linked establishments would likely exhibit temporally similar patterns of microbial test failures. Figure 1 offers an example to illustrate the concept. It shows the log of failed microbial tests at four establishments labeled A, B, C and D. Over time these facilities have been exposed to five different strains of Salmonella labeled U, V, X, Y and Z. We use the phenomenological factors to represent microbial testing data in the form of a bi-partite graph. In this graph, food establishments and individual serotypes of Salmonella become separate types of entities. A food facility node is linked with a serotype node if the particular type of Salmonella was actually isolated in a sample collected in that facility during the period of observation. Figure 2 presents topology of such graph constructed for data of Period 1 shown in Figure 1.

One problem which can be addressed using such data representation is that of link prediction. Given a few periods of observation (training data), can we predict what are the most probable Salmonella serotypes to occur at each plant during the next period? We solve this problem using Dynamic Social Network in Latent space method (DSNL) introduced in [7] and briefly described next.

### III. THE DSNL MODEL

Latent space approach to social network analysis was introduced in [8]. It was defined as a model which associates each network entity with a location in a p-dimensional Euclidean latent space, so that entities close in the latent space are more probable to form a link in the network. Our Dynamic Social Network in Latent space model (DSNL) generalized that to enable modeling evolution of relationships between entities by allowing their latent positions to change smoothly over time. The entities can move in the latent space between consecutive discrete time steps, but large moves are considered improbable [7].

Let $G_t$ be the graph observed at time step $t$. The latent position of entity $i$ at time step $t$ is denoted by the $i^{th}$ row of the $n \times k$ positions matrix $X_t$. We model two probability distributions. One generates $G_t$ from $X_t$ such that links between pairs of entities which are far away in the latent Euclidean space are less probable. The other distribution generates $X_t$ from $X_{t-1}$, and controls the smoothness of transition. These distributions also aid in tractable learning of the maximum likelihood estimates of the latent positions of the entities. We want to find $X$ that maximizes the likelihood:

$$P(X | G_t, X_{t-1}) = P(G_t | X) P(X | X_{t-1})$$

(1)

The first part of the log-likelihood form of the model equation: $\log P(G_t | X) + \log P(X | X_{t-1})$, the observation model, measures how well the latent coordinates explain the observed graph. The second part, the transition model, penalizes large changes of the latent positions from those learned in the last time step.

Let us denote the latent space distance between entities $i$ and $j$ at time step $t$ as $d_{ij}$. We also introduce radius $r_i$ of entity $i$: $r_i = \alpha \cdot \text{Degree}_i$ (the constant $\alpha$ can be learned from data). Radius captures the relative importance of an entity in the network. $r_{ij}$ equals the greater of the radii of entities $i$ and $j$. The probability of a link between entities $i$ and $j$ can then be estimated as:

$$p_{ij} = \frac{1}{1 + e(d_{ij} - r_{ij})}$$

(2)

The observation model estimates the probability that the
graph \( G_t \) was generated from coordinates \( X_t \):

\[
p(G_t|X_t) = \prod_{i \sim j} p_{ij} \prod_{i \not\sim j}(1 - p_{ij})
\]  

(3)

The observation model benefits computationally from local-ness, since only entities within radius from the entity under analysis are fully considered. It also allows a low constant probability \( \rho \) of connection between any pair of distant entities. Moreover, we use a bi-quadratic kernel to avoid quadratic computations of the observation model updates:

\[
K(d, r) = \begin{cases} 
(1 - \frac{d^2}{r^2})^2 & \text{if } d \leq r \\
0 & \text{otherwise}
\end{cases}
\]  

(4)

Therefore, the modified link probability function assumes the following final form which substantially simplifies otherwise \( O(n^2) \) computations:

\[
p_{ij}^K = Kp_{ij} + (1 - K)\rho \quad \text{if } d \leq r
\]

\[
= \rho \quad \text{otherwise}
\]  

(5)

The transition model is Gaussian:

\[
X_t \sim \mathcal{N}(X_{t-1}, \sigma^2)
\]

The parameter \( \sigma \) controls the smoothness of transition from time step \( t - 1 \) to \( t \). Large values of \( \sigma \) allow large changes of latent coordinates of the entities between subsequent time steps.

The algorithm of learning \( X_t \) has two-phases. First, the latent coordinates are initialized by a time-dependent vari-ation of multidimensional scaling (MDS) [9]. The solution combines the evidence from the current observation and the latent space locations of entities computed at the previous timestamp. These estimates are then used to initialize non-linear optimization.

Ordinary multidimensional scaling (MDS) takes as input an \( n \times n \) matrix of non-negative distances \( D \) between entities \( i \) and \( j \) in the graph. It produces an \( n \times p \) matrix \( X \) where the \( t^{th} \) row denotes the position of entity \( i \) in \( p \)-dimensional latent space. Let the coordinates of \( n \) points in a \( p \) dimensional Euclidean space be given by \( x_i, i = 1, ..., n \), where \( x_i = (x_{i1}, ..., x_{ip}) \). MDS transforms the pairwise distance matrix \( D \) into a similarity matrix \( D \) using linear transformations. The solution is given by the following formula in which \( F \) denotes Frobenius’ metric:

\[
X = \underset{X}{\text{argmin}} \|D - XX^T\|_F
\]  

(6)

In order to incorporate temporal smoothness into this objective we proposed to minimize a modified objective function:

\[
X_t = \underset{X}{\text{argmin}} \|\hat{D}_t - XX^T\|_F + \lambda \|XX^T - X_{t-1}X_{t-1}^T\|_F
\]  

(7)

The first part is identical to the standard MDS objective. The second part encourages small changes in pairwise distances between two consecutive time steps. The parameter \( \lambda \) controls the relative importance of the past and present evidence.

Conveniently, this optimization problem has a closed form solution:

\[
X_tX_t^T = \frac{1}{1 + \lambda} \hat{D}_t + \frac{\lambda}{1 + \lambda} X_{t-1}X_{t-1}^T
\]  

(8)

\( X_t \) can be obtained via eigen-decomposition of the right hand side of (8). It is possible to compute it using an iterative solver in \( O(n^2f + n + pn) \) time per iteration, where \( n \) is the number of entities, \( p \) is the number of latent dimensions, and \( f \) is the fraction of non-zero entries in the underlying matrix.

The solution (8) becomes the starting point of a nonlinear optimization for the next time step using conjugate gradient. Due to the use of the bi-quadratic kernel, the computation of the gradient of the likelihood only needs to consider entities that lie within one another’s radius. This eliminates the need for iterating over all pairs of entities and the computations can be further sped up using KD-trees [10] in \( O(rn + n \log n) \) time, where \( r \) is the average number of neighboring entities within radius.

IV. EXPERIMENTS

A trained DSNL model can be asked to predict probability of a linkage between any \{Establishment, Serotype\} pair during period \( t+1 \) based on links observed until period \( t \). To measure the ability of the model to predict future links we split the available data into independent training and testing subsets in order to objectively measure prediction accuracies. We then train the model using data from periods preceding \( t + 1 \) and we use it to separately predict the likelihood of occurrence of each serotype at each establishment during \( t + 1 \). Then, we sort the results by serotype according to the decreasing probabilities, separately for each establishment. Next, we measure recall for establishment \( A \) as the percentage of the links actually observed during period \( t + 1 \) which make it to the top \( k\% \) of the sorted results computed for \( A \) (we call \( k\% \) a recall threshold). Finally, we compute average recall and standard error statistic by aggregating results across all establishments.

We also compare DSNL results against a baseline which only considers histories of individual establishments, inde-pendently of each other. The baseline model is simple, but empirically it is quite powerful and hard to beat. For each \{Establishment, Serotype\} pair it checks the training data for their co-occurrences at subsequent periods of historical data. Then, it estimates the likelihood of future co-occurrences as an exponentially down weighted sum:

\[
l_{ij}^{\text{Base}} = \sum_{\tau=1}^{t} n_{ij(\tau)}e^{-\beta(t-\tau)}
\]  

(9)

Here, \( n_{ij(\tau)} \) denotes the number of co-occurrences of serotype \( j \) at establishment \( i \) during period \( \tau \). In our experiments we set \( \beta = 1 \), and in fact we binarize \( n_{ij(\tau)} \) so that it
is set to 1 whenever we observe at least one co-occurrence, otherwise it is set to 0.

Note that the baseline model ignores any transitivity captured by the graph structure that is being exploited using DSIGNL. It therefore maintains the assumption of mutual independence of establishments. Any gain of the predictive accuracy of the proposed approach over the baseline would indicate that relaxing the assumption of independence is in fact beneficial, and that there is predictive value in modeling establishments and serotypes as a network of entities.

The data used in our experiments is a fragment of the record of regulatory sampling of food for Salmonella conducted at a subset of establishments regulated by the U.S. Department of Agriculture (USDA) from January 2005 till December 2007. Each record in this data represents a positive result of a microbial test of a sample of food taken at a specific establishment. The data includes the information of the specific serotype of the isolated pathogen. It consists of over 7,000 records of positive tests involving about 750 unique establishments and over 90 unique serotypes.

In our previous work we have shown the utility of exploiting the structure of connectivity between food establishments and strains of Salmonella. In [6], the DSIGNL model was trained using data arranged in two observation periods corresponding to years 2005 and 2006. The trained model was then used to predict links between establishments and serotypes over the period of 2007. The predictions were then compared with the actual observations recorded in the test data. On the average, in each year the data included around 400 unique establishments, 70 serotype entities, and 1,000 links between the establishments and the serotypes.

The objective of our current work is to investigate the effects of shortening the prediction horizon. On one hand, the narrower the prediction window, the lower the expected count of occurrences of specific microbial findings per establishment, hence the greater sparseness of data per each training time step and the harder it may be to obtain reliable predictions. On the other hand, the more frequent prediction cycles, the more agile the monitoring process can be.

Figure 3 presents the recall scores computed for thresholds of \( k = 5\% \), 10\%, 15\%, 20\%, 25\% and 25\%, respectively for the DSIGNL model and the baseline. Both models use two periods of training data, each 12-month long (2005 and 2006), and also 1-year long testing period (2007). It is clear that for values of \( k \) greater than 5\%, DSIGNL algorithm outperforms the baseline. This result has already been reported in [6].

Figure 4 presents the analogical characteristic obtained when using eight non-overlapping quarterly periods to train the models. They also covered 2005 and 2006. For testing, we used the data from the first quarter of 2007. Here we observe something counter intuitive. The baseline slightly but clearly beats the DSIGNL model for recall thresholds up to 20\%. This means that when it comes to ranking of the most likely serotypes to occur at specific establishments, the model which does not take into account phenomenological hints about linkages between facilities, performs better than the one relieving the assumption of independence. At the first sight, that result may seem awkward. But, perhaps the effect is due to the increased sparseness of the graphs constructed on the basis of quarterly instead of annual aggregation of co-occurrences.

Figure 5 presents the effects of shrinking the widths of time windows of observation of density of connections between entities of the graph. We can see that the average degree of establishment entities reduces to about 72\% and the corresponding statistic for serotype entities goes down to 41\% at 3 months with respect to the degrees observed in the graphs constructed for 12-month observation windows. The relative metrics of connectivity drop further to 63\% and 27\%, respectively, for 1 month periods.

A simple explanation would be that the overly sparse graphs actually do not convey enough useful information for the DSIGNL model to prosper. A less obvious alternative would claim that there is something detrimental to the parts of data when they are provided to the algorithm in a greater
number of smaller chunks. One hypothesis that warrants verification is that the link data ages over time, and at shorter periods of observation and prediction we should pay more attention to the more recent chunks of training data. Figure IV presents results of such a test. The vertical axis denotes recall values computed for 20%-ile recall threshold at 3 month observation window width resolution. The horizontal axis denotes the amount of training data supplied to DSNL and baseline algorithms. That amount is measured with the number of the most recent 3-month chunks supplied, so that "1" corresponds to using only the most recent quarter of historical data for training, while e.g. "3" corresponds to using three 3-month chunks covering the most recent 9 months. The result supports the hypothesis stated above. The more older data we supply to train the models, the lesser the benefit from using the graph-based representation over the one that assumes independence of entities. Evidently, the microbial testing data ages relatively quickly, and patterns of relationships between establishments observed more than few quarters ago do not match the corresponding patterns of nowadays.

It is well known that prevalence of certain pathogens may be seasonal. In fact, Salmonella is known to occur more frequently in late summer and early fall than it would in winter or early spring. If such effect existed in our data, it could have a detrimental effect on the observed results, exacerbating the observed above issues with data age. Figure 7 shows another view of the recall characteristics obtained for 3-month observation periods and 20%-ile recall threshold. Here, the training data for each of the quarters of 2007 with starting dates shown on the horizontal axis, have been extracted from only the analogical quarters of 2005 and 2006. We can clearly see that marginal effects of training from only the corresponding periods of time in the past are beneficial.

V. CONCLUSION

In this paper we evaluated trade-offs between agility and reliability of predictions arising due to sparseness of data in the context of analysis of dynamic social networks. The presented work builds upon our prior research into modeling occurrences of positive results of microbial tests of food conducted at factories inspected by the U.S. Department of Agriculture.

In our model, various strains of Salmonella and food facilities are entities in a social network which evolves over time. A computationally efficient latent space model is then trained from the sequence of temporally ordered networks to predict the future links between entities. Social network based representation leads, in general, to increased predictive power when compared against alternative models that assume mutual independence of entities. The proposed approach exploits easy to observe phenomenological similarities in the past performance of food establishments.

However, the positive effects tend to diminish as the width of temporal window of analysis decreases. On one hand, the narrower the windows, the lower the expected count of
occurrences of specific microbial findings per establishment, and hence the greater sparseness of data per each training time step and the harder it is to obtain reliable predictions. On the other hand, the more frequent prediction cycles, the more agile the monitoring process can be, benefiting practical utility.

We observed that the relevance of *Salmonella* testing data diminishes over time. This effect, in some cases, renders the social network based models less predictively accurate than the simple alternative which uses assumption of independence of entities in the graph. Apparently, the relationships between establishments, measured with phenomenological factors, change over time. Facilities manifesting similar recent histories of microbial test failures are still useful in mutually boosting each others forecast. But less recent matches may be in fact detrimental to predictive accuracy. Therefore it is advisable to focus on the most recent chunks of training data when the periods of aggregation/observation become short and, therefore, when the graphs become sparse.

Domain knowledge can be of help. Models exploiting seasonality of frequency of positive tests for *Salmonella* fare better than baselines built from the same data.

The specific application domain we chose to illustrate the presented concepts is very important because *Salmonella* is among the most common causes of food borne illnesses which may result from consumption of contaminated products. The analyses similar to discussed in this paper can be used to predict occurrences of particular microbial positives at specific food factories. They could potentially aid in proactive monitoring of establishments at risk, allowing for early intervention and mitigation of adverse consequences of microbial contamination of food to public health. Additionally, the presented results can be useful in many other application domains, wherever a relatively sparse event data can be represented as links between entities forming social networks that evolve over time.

**ACKNOWLEDGMENT**

This work was partially supported by the U.S. Department of Agriculture, Centers of Disease Control and Prevention (award R01-PH000028), and the National Science Foundation under grant IIS-0325581.

**REFERENCES**


