Bird Flu Outbreak Prediction via Satellite Tracking

Yuanchun Zhou and Mingjie Tang, *Chinese Academy of Sciences*

Weike Pan, *Hong Kong Baptist University*

Jinyan Li, *University of Technology, Sydney*

Weihang Wang, Jing Shao, Liang Wu, and Jianhui Li, *Chinese Academy of Sciences*

Qiang Yang, *Huawei Noah's Ark Lab*

Baoping Yan, *Chinese Academy of Sciences*

Converting wild bird migratory paths into graphs helps achieve H5N1 outbreak prediction. A mining algorithm discovers weighted closed cliques in the graphs, and a learning algorithm then predicts potential H5N1 outbreaks.

Avian bird flu has been an ongoing topic of international concern. Here, we transform the bird-migration data analysis problem into a high-weight closed clique mining problem, and we propose a novel, High-wEight cLosed-cliquE miNing (HELEN) algorithm, which our prediction algorithm HELEN-p then uses for accurate H5N1 outbreak prediction.

Background

The H5N1 virus outbreaks in poultry in 2003, 2004, and 2009 had an unprecedented geographical impact in Asia.[1–3] The H5N1 virus is a highly pathogenic avian influenza that has emerged in southern China in the mid-1990s. A large number of wild birds died as a result of the highly pathogenic virus in Qinghai Lake, China, in 2005. The number of protected bar-headed geese had decreased 5 to 10 percent worldwide due to the epizootic disease, as estimated in 2009.[4]

The spread of H5N1 is believed to be closely related to wild-bird migration across the globe. However, effective tracking systems and data analysis tools have been lacking for a long time in China. The study on the relationship between the spread of the H5N1 virus and the bird-migration network wasn't conducted on a large scale. This situation is greatly improved now; we've collected about 1 million migration records from March 2007 to December 2009 by using a satellite tracking system and special GPS devices that ecologists attached to birds (see Figure 1). The GPS devices continuously transmitted tracking signals to the satellite, and the US Geological Survey processing unit distributed the data to researchers.

Biologists found that bird migration routes in a small area are best viewed as graph patterns like cliques[5] rather than simple

1541-1672/14/$31.00 © 2014 IEEE
Published by the IEEE Computer Society

location sequences on a small scale. It's therefore important to understand the role that migratory birds play in the ecology and transmission patterns of H5N1 by integrating data on habitats, seasonal movement chronology, routes, dates, and locations of H5N1 outbreak events. Recently, several studies have shown that H5N1 viruses in Qinghai Lake spread with the bird migration patterns.[4] Most of these analyses were conducted at a relatively coarse level of granularity (for example, between countries) and the methods for discovering the correlations of bird migration routes have limited predictive power.

Here, we mine the bird-movement pattern data and learn the relationship between graphical clique patterns and virus propagation. In particular, we use vertex weights to evaluate the seriousness of H5N1 virus transmission. Weights are differently defined by using the degree of a habitat or vertex (the frequency that birds fly among habitats), the time that birds stay at a certain habitat, or the density of the birds in a particular habitat. These weighted graph features can make the virus prediction model more accurate because we can use them to better estimate the correlations among the habitats. As a result, our prediction algorithm HELEN-p can help accurately predict future H5N1 outbreak from the migration graphs.

In our previous work, we analyzed bird virus outbreaks via mining bird migration data such as sequence rule[3] and subgraph mining.[6] In this article, we focus on how to predict future possible bird virus outbreak locations with machine learning methods. Specifically, our prediction method is based on mined high-weight closed cliques,[6] some newly developed habitat correlation criteria, and two machine learning algorithms (k-nearest neighbor, or kNN, and Laplacian-based regularized least-square, or LapRLS[7]). More importantly, with LapRLS we generalized the idea of label propagation in manifold-based, semisupervised learning to H5N1 spreads in the bird migration network.

Figure 1. A GPS tracking device attached to a bird. Ecologists captured and attached devices to the birds to monitor and analyze their migration.

Algorithm

In this section, we first introduce the basic concepts and principles of weighted graphs, and then describe the high-weight closed clique mining and H5N1 virus outbreak prediction algorithms.

Mining High-Weight Closed Cliques

In our graph-based model, a bird habitat is denoted by a node (vertex) and a migration route is denoted by an edge. A clique C is a graph with fully connected edges. If a graph G contains a clique C, then G is said to be a support graph of C. For example, graph G_1 in Figure 2 is a support graph of clique C_1 = "*abc*."

Definition 1. The frequency-support of a clique C is defined as the ratio of the number of support graphs over the total number of graphs in a database $\mathcal{D}$,

$$\text{support}^f(C) = \frac{\sum_{G \in \mathcal{D}} I(C \subseteq G)}{|\mathcal{D}|},$$

where $\sum_{G \in \mathcal{D}} I(C \subseteq G)$ is the number of support graphs of clique C, and $|\mathcal{D}|$ is the number of graphs in the database.

Given a support threshold θ^f, a clique C is a frequent clique if $\text{support}^f(C) \geq \theta^f$. In addition, if there doesn't exist another clique C' satisfying $C \subseteq C'$ and $\text{support}^f(C') = \text{support}^f(C)$, C is a frequent closed clique. Closed cliques are important since they greatly reduce the number of child cliques with the same support level. Frequent-closed-clique mining finds all frequent closed cliques from a graph database. Given the graph database in Figure 2 and $\theta^f = 0.5$, "*abc*" and "*abde*" are two frequent and closed cliques.

The weight of a vertex v is denoted by weight(v). We consider three weighting ideas in this work:

- $W_{\text{frequency}}$, which measures how frequently a bird flies among different habitats;
- $W_{\text{time}} = t_{\text{arrive}} - t_{\text{leave}}$, which measures how long a bird stays at a certain habitat, where t_{arrive} and

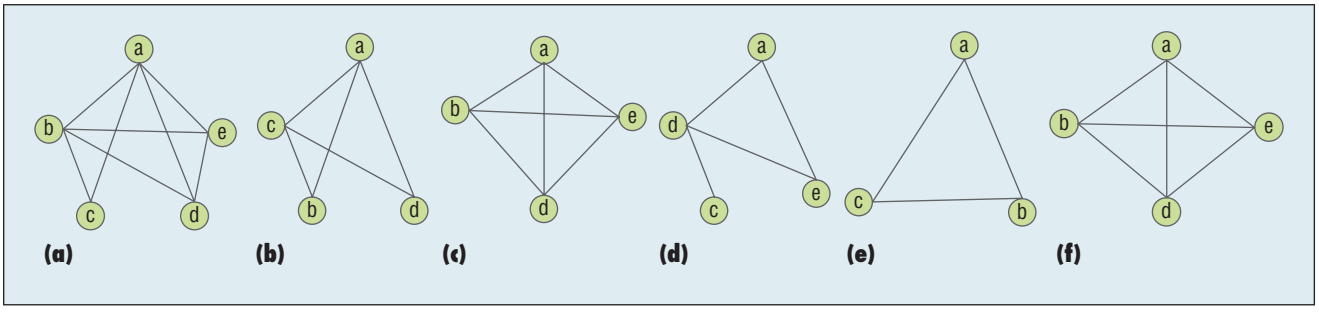

Figure 2. A graph database (weight(a) = 7, weight(b) = 6, weight(c) = 2, weight(d) = 14, weight(e) = 20). Then, weight of cliques "abc" and "abde" are 15 and 47, respectively. (a) Graph G_1. (b) graph G_2. (c) Graph G_3. (d) Graph G_4. (e) Clique "*abc*." (f) Clique "*abde*."

t_{leave} are the bird's arrival and departure times; and

- $W_{density}$, which measures bird density in the habitat, and is calculated by using the habitat's area size divided by the number of migration records received by the satellite tracking system from that habitat.

The weight of a graph G is given by $\text{weight}(G) = \Sigma_{v \in G} \text{weight}(v)$.

Definition 2. The weight-support of a clique C is defined as

$$\text{support}^w(C) = \frac{\text{weight}(C)\sum_{G\in\mathcal{D}} I(C \subseteq G)}{\sum_{G\in\mathcal{D}} \text{weight}(G)}, \quad (1)$$

where the numerator $\text{weight}(C)\sum_{G\in\mathcal{D}} I(C \subseteq G)$ denotes the total weight of the clique C in database $\mathcal{D}$, and the denominator $\sum_{G\in\mathcal{D}} \text{weight}(G)$ is simply a normalization term. Given a support threshold θ^w, a clique C is a high-weight-support clique if $\text{support}^w(C) \geq \theta^w$. In addition, if no other clique C' exists that satisfies $C \subseteq C'$ and $\text{support}^w(C') \geq \text{support}^w(C)$, then C is a high-weight-support closed clique (HWCC). We wish to find all frequent and closed cliques from graph database $\mathcal{D}$ with respect to the vertex weight. For example, given the graph database in Figure 2, we have $\text{support}^w(\text{"}abc\text{"}) = (15 \times 2)/(49 + 29 + 47 + 43) = 0.18$, $\text{support}^w(\text{"}abde\text{"}) = (47 \times 2)/(49 + 29 + 47 + 43) = 0.56$. If $\theta^w = 0.5$, the clique "*abde*" is a high-weight closed clique.

Definition 3. The graph-weight-support of a clique C is defined as

$$\text{support}^g(C) = \frac{\sum_{G\in\mathcal{D}} I(C \subseteq G)\,\text{weight}(G)}{\sum_{G\in\mathcal{D}} \text{weight}(G))}, \quad (2)$$

where the numerator $\sum_{G\in\mathcal{D}} I(C \subseteq G)\,\text{weight}(G)$ denotes the total weight of support graphs of clique C in database $\mathcal{D}$, and the denominator $\sum_{G\in\mathcal{D}} \text{weight}(G)$ is again for normalization. Given a support threshold θ^g, a clique C is a high-graph-weight-support clique if $\text{support}^g(C) \geq \theta^g$. In addition, if there doesn't exist a clique C' satisfying $C \subseteq C'$ and $\text{support}^g(C') = \text{support}^g(C)$, C is a high-graph-weight-support closed clique (HGWCC).

The downward-closure property (or anti-monotone property), which has been widely used to accelerate pattern-mining algorithms, states that any child pattern (for example, a subset of vertices) of a frequent pattern is also frequent. Hence, if no $k-1$-patterns are frequent, we don't need to explore k-patterns. However, we observe that the downward-closure property doesn't hold true in HWCC mining. For example, in Figure 2, $\text{support}^w(\text{"}abde\text{"}) = 0.56$, $\text{support}^w(\text{"}abd\text{"}) = 0.32$. If we set the support threshold $\theta^w = 0.5$, then "*abd*" is a low-weight clique, while its parent -graph "*abde*" is a high-weight clique. So, this causes difficulties for mining algorithms. If it can be proved that if any $k-1$-clique $C^{[k-1]}$ isn't a high-graph-weight-support clique, then k-clique $C^{[k]}$ isn't either. This downward-closure property is useful in the process of enumerating cliques. If we know that a $k-1$-clique $C^{[k-1]}$ isn't a high-graph-weight-support clique, there's no need to enumerate any k-clique. It can be also proved that if $\theta^w = \theta^g$, then HWCC $\subseteq$ HGWCC.

The main idea of the HELEN algorithm is to search over a clique lattice, as Figure 3 shows. The algorithm's pseudocodes cover three major computational steps.

Input: Graph database $\mathcal{D}$ and vertex weight, threshold θ^g and θ^w;
Output: HWCC.

1. Calculate the graph weight using $\mathcal{D}$ and vertex weight.
2. Search the lattice and obtain HGWCC using $\mathcal{D}$, vertex weight, and θ^g.
3. Check the HGWCC and obtain HWCC using $\mathcal{D}$, vertex weight, and θ^w.

The mined HWCCs from the illustration data are marked with red circles in Figure 3.

Calculating Habitat Correlation

Our prediction method also involves two types of habitat correlations: location- and clique-based.

Definition 4. For any two habitats i and j, the location-based correlation is defined by the distance

d_{ij} of the two habitats, calculated using

$$\frac{1/d_{ij}}{\max_{ij} 1/d_{ij}}, \quad (3)$$

where the denominator, $\max_{ij} 1/d_{ij}$, is a normalization term to make the correlation in the range of [0, 1].

We consider two types of distance in our correlation estimation:

- The Euclidean distance $d_{ij}^{ec} = \sqrt{(\phi_i - \phi_j)^2 + (\lambda_i - \lambda_j)^2}$, where (ϕ_i, λ_i) and (ϕ_j, λ_j) are the latitude and longitude of habitats i and j, respectively.
- The great-circle distance[8] $d_{ij}^{gc} = r\Delta\hat{\sigma}_{ij}$, where r is the radius $\Delta\lambda = \lambda_i - \lambda_j$, and

$$\Delta\hat{\sigma}_{ij} = \arctan\left(\frac{\sqrt{(\cos\phi_j \sin\Delta\lambda)^2 + (\cos\phi_i \sin\phi_j - \sin\phi_i \cos\phi_j \cos\Delta\lambda)^2}}{\sin\phi_i \sin\phi_j + \cos\phi_i \cos\phi_j \cos\Delta\lambda}\right).$$

Definition 5. For any two habitats i and j, the clique-based correlation is defined by using the weighted supports of closed cliques to which i and j belong:

$$c_{ij}^{w} = \frac{\sum_{C \in \mathcal{C}} I\big((i,j) \subseteq C\big)\,\text{support}^{w}(C)}{\max_{ij} \sum_{C \in \mathcal{C}} I\big((i,j) \in C\big)\,\text{support}^{w}(C)}, \quad (4)$$

where $\mathcal{C}$ is a set of HWCCs, and $\sum_{C \in \mathcal{C}} I\big((i,j) \subseteq C\big)\,\text{support}^{w}(C)$ denotes the summation of the weighted support of the closed cliques to which the habitats i and j belong.

For example, in Figure 3, C = {"abde," "ad," "ade"} and $\sum_{C \in \mathcal{C}} I\big((a,e) \subseteq C\big)\,\text{support}^{w}(C) = \text{support}^{w}(\text{"}abde\text{"}) + \text{support}^{w}(\text{"}ade\text{"})$. The correlations among "a," "b," "c," "d," and "e" are as follows: $c_{ab}^{w} = 0.31$, $c_{ac}^{w} = 0$, $c_{ad}^{w} = 1$, $c_{ae}^{w} = 0.72$, $c_{bc}^{w} = 0$, $c_{bd}^{w} = 0.31$, $c_{be}^{w} = 0.31$, $c_{cd}^{w} = 0$, $c_{ce}^{w} = 0$, and $c_{de}^{w} = 0.72$.

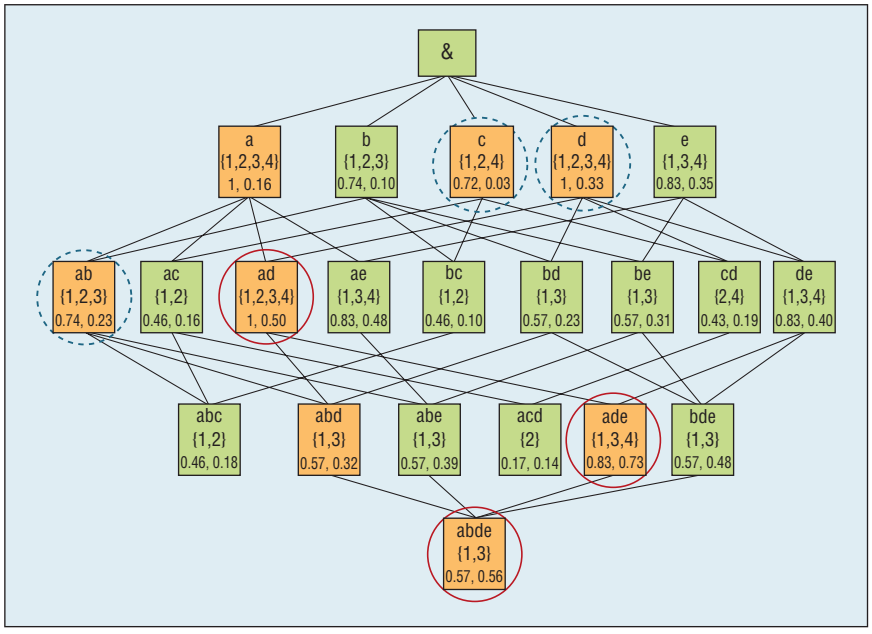

Figure 3. A clique lattice with the graphs from Figure 2. Each rectangle contains a clique (for example, "*ab*"), a corresponding set of graphs to which the clique belongs (for example, {1, 2, 3}), a graph-weight-support (for example, support^{g}(*C*) = 0.74 via Equation 2), and a weight-support (for example, support^{w}(*C*) = 0.23 via Equation 1). The rectangles in yellow denote the depth-first search space with $\theta^{g} = 0.5$, and the search order is "a," "ab," "abd," "abde," "ad," "ade," "c," and "d." The rectangles with circles are high-graph-weight-support closed cliques with $\theta^{g} = 0.5$, among which the rectangles with solid red circles are the final high-weight-support closed cliques with $\theta^{w} = 0.5$.

The Prediction Algorithm

We take the following pseudocodes in the prediction of H5N1 virus outbreaks:

Input: Graph database $\mathcal{D}$, vertex weight, threshold θ^{g} and θ^{w}, positive instance p, number of predicted habitats k;

Output: A ranked list of k predicted habitats.

1. Call the HELEN algorithm to obtain HWCC.
2. Calculate the correlations of any two habitats according to Equations 3 or 4 using the mined HWCC.
3. Run the kNN or LapRLS algorithm to find the top k likely outbreak habitats.

The problem setting of our prediction task is transductive learning rather than inductive learning, where the input includes one positive instance (that is, labeled training data), many unlabeled instances (that is, unlabeled test data), and correlations among the instances. A common supervised machine learning method trains a prediction model using labeled training data only, for which one single positive instance (that is, training data in our problem) isn't sufficient.

The two machine learning methods kNN and LapRLS are explained as follows. We hypothesize that a H5N1 outbreak is highly correlated with the migration network, which is reflected in the mined high-weight closed cliques. We verify this hypothesis in the experimental section later in the article. Given a habitat with an H5N1 outbreak (Habitat$_p$) and the habitat correlation (c_{ip}^{ec}, c_{ip}^{gc}, or c_{ip}^{w}), we can rank the remaining habitats

Table 1. Description of the data used in the experiments.

Bird type	Bird number	Active time		Stay (days)		Migration record number	H5N1 rate (%) and number of birds confirmed using reverse transcription-polymerase chain reaction
		Start	End	Max	Min		
Bar-headed geese	29	21 March 2007	21 October 2009	745	48	783,240	2.27 12 of 528
Ruddy shelduck	20	21 March 2007	1 February 2009	347	28	179,302	2.17 3 of 138
Brown-headed gull	10	21 June 2007	7 June 2008	159	41	37,242	3.60 14 of 389

and obtain the top k habitats with the largest correlation based on the kNN method. For example, if "a" in Figure 3 is taken as a positive habitat, we have the ranking list of "d," "e," "b," and "c" according to the correlations. We denote the corresponding HELEN-p variant as HELEN-p(kNN).

Under a kernel-learning approach, we take the originating habitat of the H5N1 outbreak as a single positive instance. We predict other outbreak habitats by using the LapRLS method, where the normalized Laplacian matrix $\mathcal{L}$ is calculated based on a habitat correlation matrix $\mathbf{W} = \left[c_{ij}^{w}\right] \in \mathrm{R}^{n \times n}$,

$$\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2},$$
where $\mathbf{D} = \mathrm{diag}(\mathbf{W}\mathbf{1})$,

where $\mathbf{I}$ is an identity matrix and $\mathbf{1}$ is the vector with all entry values of 1.

Then, we apply the LapRLS objective function with a single positive instance,

$$\min_{f} \boldsymbol{f}'\mathrm{L}\boldsymbol{f} + \frac{\alpha}{n}\|\boldsymbol{f} - \boldsymbol{y}\|_F^2,$$

where $\boldsymbol{f} \in \mathbb{R}^{n \times 1}$ is the prediction vector, $\boldsymbol{y}$ is the label vector with $\boldsymbol{y}_i = \begin{cases} 1 \text{ if } i = p, \\ 0 \text{ if } i \neq p. \end{cases}$, $\|\cdot\|_F$ denotes the Frobenius norm, and α is the tradeoff parameter. Hence, the final obtained score vector $\boldsymbol{f}$ can be used to rank the remaining habitats and find the top k habitats with the highest probability of an H5N1 outbreak. We denote the corresponding HELEN-p variant as HELEN-p(LapRLS).

Compared with the HELEN-p(kNN) method, HELEN-p(LapRLS) has the potential of bridging two habitats beyond k-nearest neighbors, because it can propagate the label via local connections,[7,9] which is also supported by our experimental results.

Experiments

In this section, we first describe the real-world bird migration data and then show our mining and prediction results.

Data Collection

We conducted our on-site studies at the Qinghai Lake National Nature Reserve, Qinghai Province, China, between March 2007 and December 2009. Ecologists randomly captured 59 birds from different flocks and tied a battery-powered GPS device to each of them. Table 1 presents more details of the data. We collected nearly 1 million migration records of the 59 birds by 25 December 2009. We selected 29 bar-headed geese for our subsequent analysis of the same type of birds. The 29 bar-headed geese correspond to 29 graphs (one for each bird) in our algorithms, and each graph contains the same 103 nodes corresponding to 103 habitats.

We used the reverse transcription-polymerase chain reaction technique (see www.who.int/influenza/resources/documents/RecAIlabtestsAug07.pdf) to confirm whether a bird is or isn't infected with the virus, and hence to determine the prevalence of H5N1 in Qinghai Lake. We tested 1,055 samples (birds). The experiments confirmed that 12 bar-headed geese, three ruddy shelducks, and 14 brown-headed gulls are positive for an H5N1 subtype. These data are compared to the total numbers of birds of the three types (see the last column of Table 1), and it can be seen that the prevalence of H5N1 in Qinghai Lake was high. To obtain the relationship between migratory birds and H5N1 outbreaks, we extracted information about H5N1 outbreaks from the Ministry of Agriculture of the People's Republic of China Database and the World Organization for Animal Health (OIE) Database for the period of February 2004 to May 2009.

Summary of Experimental Results

We conduct empirical studies of H5N1 outbreak analysis and prediction using the mined cliques in the following two subsections.

H5N1 outbreak analysis using mined cliques. We applied the HELEN algorithm to those 29 graphs to extract cliques. Figure 4 shows one high-weight clique C_{15}. If we only consider its frequency support (support^f= 3/29), C_{15} would be pruned. However, this clique has a weight of 0.13, 0.16, and 0.052, respectively, according to $W_{\text{frequency}}$, W_{time}, and W_{density} weighting strategies, and it contributes to more than 5.2 percent of the total time of the birds' spring migration time. Table 2 shows that the migration network has a strong relationship with H5N1 outbreaks. For example, while birds prefer to stay at habitat 4 (H_4), three

cases of H5N1 outbreak are reported. In addition, this clique shows that the habitat H_4 has a strong correlation with its neighboring habitats (H_1, H_2, H_3, and H_5) under the high weight of $W_{density}$. Interestingly, habitats (H_2, H_3, and H_5) are also reported to have H5N1 outbreaks. The weight of those habitats does reflect the possibility of virus transmission.

A total of 24 percent of our mined cliques have a low frequency but a high weighted support. This magnifies the importance of weight clique mining, because otherwise, these low-frequency cliques would be pruned by the traditional frequent-closed-clique mining algorithms. High-weight closed-clique mining can help biological professionals make better decisions, for example, by pointing out some high-weighted cliques. More mining results can be found at www.qinghailake.csdb.cn/qhlakesdm/page/paper/link1.htm.

H5N1 outbreak prediction using mined cliques. Our algorithm mined 245 cliques from the 29 graphs and 103 habitats ($\theta^g = 0$, $\theta^w = 0$), where each clique has four different weights $W_{frequency}$, W_{time}, $W_{density}$, and support^f. Among those 103 habitats, 16 had one or more cases of H5N1 outbreaks—that is, they're positive habitats. In each prediction test, we take one positive habitat out of those 16 habitats, and report the averaged results over the 16 times. To gain more insights on HWCC and the effect of the support threshold θ^w, we first study the prediction performance when $\theta^w = 0$, and then increase its value gradually to 0.05, 0.1, and 0.15.

Table 3 shows the prediction results when $\theta^w = 0$. We can see two important points: the approach of using clique-based correlation is much better than that of using the habitats' geometric information, confirming the usefulness of the bird satellite tracking system or migration network in habitat correlation estimation; and although the clique-based correlation might fail to build connections between two habitats that never appear in any of the same cliques, as shown by the results of HELEP-p(kNN), HELEN-p(LapRLS) can complement this weakness via label propagation (or H5N1 spread). More empirical studies of HELEN-p(kNN) and HELEN-p(LapRLS) can be found at www.qinghailake.csdb.cn/qhlakesdm/page/paper/link3.htm, from which we can see that HELEN-p(LapRLS) improves the prediction performance and beats kNN in all cases.

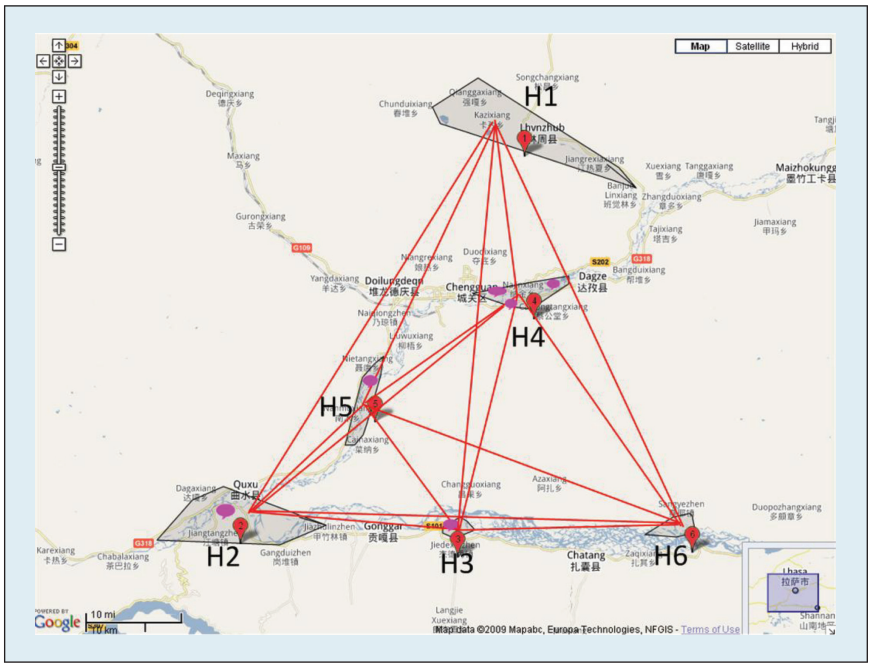

Figure 4. A mined high-weight closed clique, C_{15}, with low frequency support (support^f = 3/29).

Table 2. Detailed information of the habitats and weight of clique C_{15}.

Habitat	$W_{frequency}$	W_{time}	$W_{density}$	Outbreak cases
H_1	18	100	800	N/A
H_2	34	140	130	1
H_3	35	109	103	1
H_4	31	173	270	3
H_5	48	9	69	1
H_6	24	19	78	N/A

Figure 5 shows the prediction performance of HELEN-p(LapRLS) with different values of θ^w. We can see that using a relatively larger threshold improves prediction performance in most cases. This observation can be explained by the fact that a reduction of noise in the clique weights can result in a better correlation estimation in Equation 4. However, using a too-large threshold could reduce the prediction performance, which makes sense because the correlation between two habitats might not appear when using too-few selected closed cliques. Therefore, we can conclude that using a relatively higher threshold is better

Table 3. The H5N1 outbreak prediction performance of the HELEN-p using habitat correlation estimated from geometric locations and migration data of the bird satellite tracking system.*

Evaluation metric	Geometric locations		Using bird satellite tracking system							
	HELEN-p(*k*NN)		HELEN-p(*k*NN), c_{ij}^{w}				HELEN-p(LapRLS), c_{ij}^{w}			
	c_{ij}^{gc}	c_{ij}^{ec}	$W_{frequency}$	W_{time}	$W_{density}$	support^{f}	$W_{frequency}$	W_{time}	$W_{density}$	support^{f}
Pre@1	$0.13_{\pm 0.34}$	$0.31_{\pm 0.48}$	$0.63_{\pm 0.50}$	$0.56_{\pm 0.51}$	$0.63_{\pm 0.50}$	$0.63_{\pm 0.50}$	$0.88_{\pm 0.34}$	$1_{\pm 0}$	$0.94_{\pm 0.25}$	$0.88_{\pm 0.34}$
Pre@5	$0.10_{\pm 0.13}$	$0.20_{\pm 0.18}$	$0.58_{\pm 0.28}$	$0.56_{\pm 0.26}$	$0.56_{\pm 0.28}$	$0.56_{\pm 0.23}$	$\mathbf{0.85}_{\pm 0.27}$	$0.76_{\pm 0.08}$	$0.84_{\pm 0.13}$	$\mathbf{0.85}_{\pm 0.15}$
Pre@10	$0.15_{\pm 0.09}$	$0.15_{\pm 0.12}$	$0.44_{\pm 0.13}$	$0.45_{\pm 0.12}$	$0.45_{\pm 0.12}$	$0.44_{\pm 0.13}$	$\mathbf{0.57}_{\pm 0.05}$	$0.55_{\pm 0.05}$	$0.55_{\pm 0.05}$	$0.56_{\pm 0.05}$
Pre@15	$0.14_{\pm 0.08}$	$0.14_{\pm 0.08}$	$0.37_{\pm 0.09}$	$0.38_{\pm 0.09}$	$0.37_{\pm 0.09}$	$0.35_{\pm 0.08}$	$\mathbf{0.50}_{\pm 0.04}$	$0.42_{\pm 0.03}$	$0.42_{\pm 0.03}$	$0.48_{\pm 0.04}$

* Note that Pre@k = $\frac{\#\text{positive habitat}}{k}$, threshold $\theta^{w} = 0$ and $\alpha = 1$.

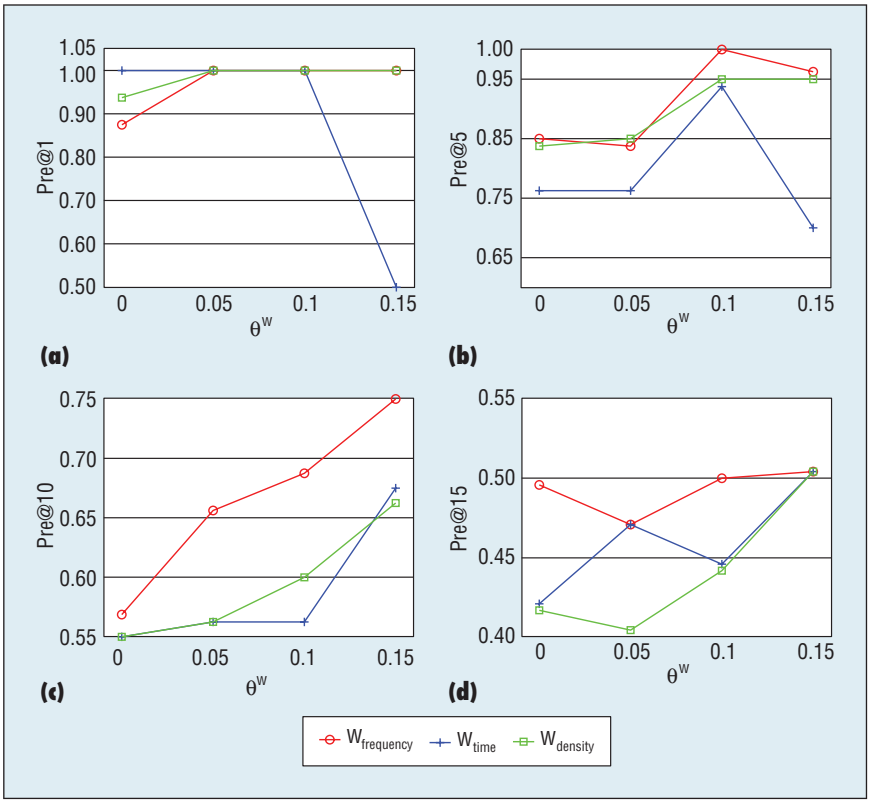

Figure 5. The H5N1 outbreak prediction performance of HELEN-p(LapRLS) with different values of θ^{w} on evaluation metrics (a) Pre@1, (b) Pre@5, (c) Pre@10, and (d) Pre@15.

in prediction, which supports our assumption that H5N1 spreads via high-weight closed cliques.

In this article, we've developed a novel H5N1 outbreak prediction algorithm (HELEN-p) that makes use of the mined cliques and machine learning methods. Our assumption that H5N1 spreads via high-weight closed cliques and frequent cliques is also supported by our experimental results (see www.qinghailake.csdb.cn/qhlakesdm/page/paper/link1.htm and www.qinghailake.csdb.cn/qhlakesdm/page/paper/link2.htm for more information). For future work, we'll explore more sophisticated algorithms to integrate different weighting strategies and contextual constraints.[10] Some preliminary results using linear combinations have been obtained (see www.qinghailake.csdb.cn/qhlakesdm/page/paper/link4.htm and www.qinghailake.csdb.cn/qhlakesdm/page/paper/link5.htm). ■

Acknowledgments

We thank the Natural Science Foundation of China (grants 61003138 and 91224006) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grants X-DA06010202 and XDA05050601) for their support. Qiang Yang thanks Hong Kong Research Grants Council (grants 621211 and 620812) for its support.

This paper is extended from our previous work.[6] Jianhui Li is the corresponding author for this work.

References

1. H. Chen et al., "Avian Fu: H5N1 Virus Outbreak in Migratory Waterfowl," *Nature*, vol. 436, July 2005, pp. 191–192.
2. J. Liu et al., "Highly Pathogenic H5N1 Influenza Virus Infection in Migratory Birds," *Science*, vol. 309, no. 5738, 2005, p. 1206.
3. M. Tang et al., "Exploring the Wild Birds' Migration Data for the Disease Spread Study of H5N1: A Clustering and Association Approach," *Knowledge and Information Systems*, vol. 27, May 2011, pp. 227–251.
4. Z. Kou et al., "The Survey pf H5N1 Flu Virus in Wild Birds in 14 Provinces of China from 2004 to 2007," *PLOS ONE*,

vol. 4, no. 9, 2009; doi:10.1371/journal.pone.0006926.
5. Y.-S. Hou et al., "Distribution and Diversity of Waterfowl Population in Qinghai Lake National Nature Reserve," *Acta Zootaxonomica Sinica*, vol. 34, no. 1, 2009, pp. 184–187.
6. M. Tang et al., "Birds Bring Flues? Mining Frequent and High Weighted Cliques from Birds Migration Networks," *Proc. 15th Int'l Conf. Database Systems for Advanced Applications*, vol. 2, 2010, pp. 359–369.
7. M. Belkin P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Machine Learning Research*, vol. 7, Nov. 2006, pp. 2399–2434.
8. T. Vincenty, "Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations," *Survey Review*, vol. 23, no. 176, 1975, pp. 88–93.
9. X. Ling et al., "Spectral Domain-Transfer Learning," *Proc. 14th Ann. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2008, pp. 488–496.
10. Q. Yang, "A Theory of Conflict Resolution in Planning," *Artificial Intelligence*, vol. 58, nos. 1–3, 1992, pp. 361–392.

THE AUTHORS

Yuanchun Zhou is an associate professor and director assistant of the Scientific Data Center at the Computer Network Information Center, Chinese Academy of Sciences. His research interests include bid data mining, cloud computing, and data-intensive computing. Zhou has a PhD in computer science from the Institute of Computing Technology, Chinese Academy of Sciences. Contact him at zyc@cnic.cn.

Mingjie Tang is a graduate student at Purdue University. His research interests include databases, Big Data analysis, and data mining. Tang has an MS in computer science from the Graduate University of Chinese Academy of Sciences. Contact him at mingjie@cnic.cn.

Weike Pan is a lecturer with the College of Computer Science and Software Engineering, Shenzhen University. He's also the information officer of *ACM Transactions on Intelligent Systems and Technology*. His research interests include transfer learning, recommender systems, and statistical machine learning. Pan has a PhD in computer science and engineering from the Hong Kong University of Science and Technology. Contact him at panweike@szu.edu.cn.

Jinyan Li is an associate professor and core member at Advanced Analytics Institute and Center for Health Technologies, Faculty of Engineering and IT, University of Technology, Sydney. His research interests include fundamental data mining algorithms, machine learning, gene expression data analysis, structural bioinformatics, and information theory. Jinyan has a PhD in computer science from the University of Melbourne. Contact him at Jinyan.Li@uts.edu.au.

Weihang Wang is a PhD student at Purdue University. Her research interests include distributed systems, cloud computing, and networks. Wang has an MS in computer science from the Graduate University of Chinese Academy of Sciences. Contact her at wang1315@purdue.edu.

Jing Shao is an engineer at the Computer Network Information Center, Chinese Academy of Sciences. His research interests include moving object mining, Big Data mining, and data analysis. Shao has an MS in computer science from the Graduate University of Chinese Academy of Sciences. Contact him at jingshao@cnic.cn.

Liang Wu is a master's student at the Computer Network Information Center, Chinese Academy of Sciences. His research interests include data mining and machine learning. Wu has a BS in computer science from Beijing University of Posts and Telecommunications. Contact him at wuliang@cnic.cn.

Jianhui Li is a professor at the Computer Network Information Center, Chinese Academy of Sciences. His research interests include Big Data mining, large-scale distributed database management and integration, semantic-based data integration, data-intensive computing, and scientific applications. Li has a PhD in computer science from the Institute of Computing Technology, Chinese Academy of Sciences. Contact him at lijh@cnic.cn.

Qiang Yang is the head of Huawei Noah's Ark Lab, Hong Kong. He's also a professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. His research interests include data mining and artificial intelligence. Yang has a PhD in computer science from the University of Maryland, College Park. His research teams won the 2004 and 2005 ACM KDD CUP competitions on data mining. He was the vice chair of ACM Sigart, the founding editor in chief of the *ACM Transactions on Intelligent Systems and Technology,* an AAAI Fellow, IEEE Fellow, International Association of Pattern Recognition (IAPR) Fellow, and ACM Distinguished Scientist. Contact him at qyang@cse.ust.hk.

Baoping Yan is a professor and chief engineer at the Computer Network Information Center, Chinese Academy of Sciences. Her research interests include computer network systems, industrial automation and CIMS (computer integrated manufacturing system) network technology, ATM-based workstation cluster systems, large-scale networks and system integration, and Internet/intranet comprehensive information management systems. The Chinese government has granted her a special allowance for her outstanding contributions. Contact her at ybp@cnic.cn.