H5N1 Outbreak Prediction Using A Satellite Bird-tracking System

Yuanchun Zhou and Mingjie Tang
Computer Network Information Center
Chinese Academy of Sciences
{zyc,mingjie}@cnic.cn

Weike Pan
Hong Kong University of Science and Technology
weikep@cse.ust.hk

Jinyan Li
University of Technology Sydney
Jinyan.Li@uts.edu.au

Weihang Wang, Jing Shao, Liang Wu and Jianhui Li*
Computer Network Information Center
Chinese Academy of Sciences
{jingshao,wuliang,lijh}@cnic.cn

Qiang Yang
Hong Kong University of Science and Technology
qyang@cse.ust.hk

Baoping Yan
Computer Network Information Center
Chinese Academy of Sciences
ybp@cnic.cn

Abstract

Advanced satellite tracking technologies have collected a huge amount of wild birds' migration data. These data are very useful for biologists to understand birds' dynamic migration patterns, to study correlations between the habitats, and to predict global spread trends of avian influenza. We transform the biological problem into a machine learning problem by converting the migratory paths of wild birds to graphs. Our first step of H5N1 outbreak prediction is to discover weighted closed cliques from the graphs by our mining algorithm HELEN (short for High-wEight cLosed cliquE miNing), which are then used by our learning algorithm HELEN-p to predict potential H5N1 outbreaks at habitats. We show that the prediction is more accurate in comparison with the traditional method on the migration data obtained through a real satellite bird-tracking system. It is also confirmed by our empirical analysis that H5N1 spreads in a manner of high-weight closed cliques and frequent cliques.

Keywords
Computational Sustainability; Bird Flu Prediction; Wild-Bird Migration Data Mining; H5N1 Prediction in Qinghai Lake, China; Machine Learning and Graph Mining.

[1]The corresponding author is Jianhui Li.

1 Introduction

Figure 1: A GPS tracking device attached to a bird.

The H5N1 virus outbreaks in poultry in 2003, 2004 and 2009 caused unprecedented geographical impact in Asia [2, 6, 8]. The H5N1 virus is a highly pathogenic avian influenza (HPAI) 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; and the number of the protected bar-headed geese had decreased $5\% - 10\%$ worldwide due to the epizootic disease alone as estimated in 2009 [5].

The spread of H5N1 is believed to be closely related to

wild-bird migration across the globe [4]. However, as 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 was not conducted in large scale. This situation is greatly improved now, and we have collected the movement data of about one million records from March 2007 to December 2009 by using a satellite tracking system and special GPS devices attached to birds (see Fig. 1). Specifically, migration birds were captured by ecologists and set up with GPS mobile sensor devices. And the tracking signals were then transmitted to the satellite continuously and the data were distributed by the USGS processing unit which were then received by the researchers.

Biologists found that bird migration routes in a small area can be best viewed as *graph patterns* like cliques [3] rather than simple location sequences in small scale. It is therefore important to understand the role that migratory birds play in the ecology and the transmission patterns of H5N1 by integrating data on habitats, seasonal movement chronology, routes, dates, and locations of H5N1 outbreak events. Recently, several studies at Qinghai Lake have shown that H5N1 viruses in Qinghai Lake spread with the bird migration patterns [5]. Most of these analysis were conducted at a relative coarse level of granularity (e.g. between countries) and the methods for discovering the correlations of bird migration routes have limited predictive power.

In this paper, we take a data mining and machine learning approach to exploit the collected data to build a bird-virus prediction model. We mine the bird-movement patterns, and learn the relationship between graphical clique patterns and virus propagation. In particular, we use vertex weights as an important factor to evaluate the seriousness of H5N1 virus. 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 they can be used to estimate the correlations among the habitats better. As a result, our prediction algorithm HELEN-p can be used to accurately predict the future H5N1 outbreak from the migration graphs.

Our main contributions are summarized as bellow,

1. we transform the bird-migration data analysis problem into a high-weight closed clique mining problem; and
2. we propose a novel high-weight closed-clique mining algorithm (HELEN), which is then used by our prediction algorithm HELEN-p for accurate H5N1 outbreak prediction.

Compared with our previous work, we have extended previous work significantly. In our previous work, we analyzed bird virus outbreak reasons via mining the birds migration data such as sequence rule mining [8] and sub-graph mining [7]. In this paper, we focus on how to predict the future possible bird virus outbreak locations by machine learning methods. Specifically, our prediction method is based on the mined high-weight closed cliques, some newly developed habitat correlation criteria, and two machine learning algorithms (i.e., kNN and LapRLS [1]). More importantly, in LapRLS, we generalized the idea of *label propagation* in manifold based semi-supervised learning to *H5N1 spreads* in the bird migration network.

2 Algorithm

2.1 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 Fig. 2 is a support graph of clique $C_1 =$ "abc" (Fig. 2(e)).

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* $\mathcal{D}$*,*

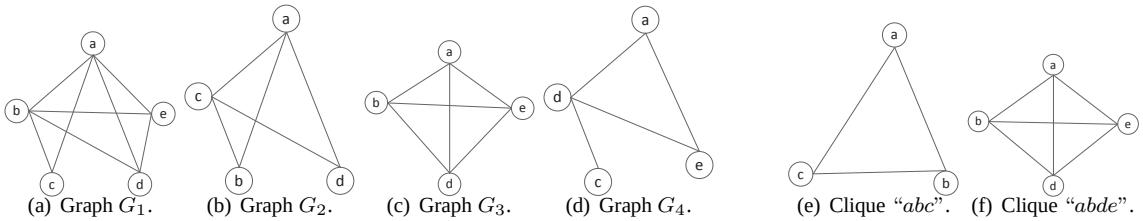
$$support^f(C) = \frac{\sum_{G \in \mathcal{D}} I(C \subseteq G)}{|\mathcal{D}|}, \quad (1)$$

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 $support^f(C) \geq \theta^f$. In addition, if there does not exist another clique C' satisfying $C \subseteq C'$ and $support^f(C') = support^f(C)$, C is a frequent closed clique (FCC). Closed cliques are important since they greatly reduce the number of child cliques with the same level of support. FCC mining is to find all frequent closed cliques from a graph database. Given the graph database in Fig. 2 and $\theta^f = 0.5$, "abc" and "$abde$" are two frequent and closed cliques.

The weight of a vertex v is denoted by $weight(v)$. Three weighting ideas are considered by this work:

1. $W_{frequency}$ (or $W_{freq.}$) which measures how frequently a bird flies among different habitats.
2. $W_{time} = t_{arrive} - t_{leave}$ which measures how long a bird stays at a certain habitat, where t_{arrive} and t_{leave} are the arrival time and departure time of the bird.
3. $W_{density}$ (or $W_{dens.}$) which measures the density of the birds in the habitat, and is calculated by using the



(a) Graph G_1. (b) Graph G_2. (c) Graph G_3. (d) Graph G_4. (e) Clique "abc". (f) Clique "$abde$".

Figure 2: A graph database ($weight(a) = 7$, $weight(b) = 6$, $weight(c) = 2$, $weight(d) = 14$, $weight(e) = 20$).

area size of the habitat to divide the number of migration records received by the satellite tracking system from the habitat.

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

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

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

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

$$support^g(C) = \frac{\sum_{G \in \mathcal{D}} I(C \subseteq G) weight(G)}{\sum_{G \in \mathcal{D}} weight(G))}, \qquad (3)$$

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

The "downward closure" property (anti-monotone property), which has been widely used to accelerate pattern mining algorithms, states that any child pattern (e.g. a subset of vertices) of a frequent pattern is also frequent. Hence, if no k-1-patterns are frequent, we do not need to explore k-patterns. However, we observe that the "downward closure" property does not hold in HWCC mining. For example, in Fig. 2, $support^w(\text{"}abde\text{"}) = 0.56$, $support^w(\text{"}abd\text{"}) = 0.32$. 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. It can be proved that if any k-1-clique $C^{[k-1]}$ is not a high-graph-weight-support clique, then k-clique $C^{[k]}$ is not either. This "downward closure" property is useful in the process of enumerating cliques. If we know that a k-1-clique, $C^{[k-1]}$ is not a high-graph-weight-support clique, there is 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 HELEN algorithm is to search over a clique lattice as shown in Fig. 3. Its pseudo codes covering three major computational steps are presented as follows:
Input: Graph database $\mathcal{D}$ and vertex weight, threshold θ^g and θ^w;
Output: HWCC.
Step 1: Calculate the graph weight using $\mathcal{D}$ and vertex weight; Step 2: Search the lattice and obtain HGWCC using $\mathcal{D}$, vertex weight and θ^g; and Step 3: Check the HGWCC and obtain HWCC using $\mathcal{D}$, vertex weight and θ^w.

The mined HWCCs from the illustration data are shown in the last column of the Table in Fig. 3.

2.2 Calculating Habitat Correlation

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

Definition 4 *For any two habitats i and j, the* location-based correlation *is defined by the distance d_{ij} of the two habitats. It is calculated by using*

$$\frac{1/d_{ij}}{\max_{ij} 1/d_{ij}}, \qquad (4)$$

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

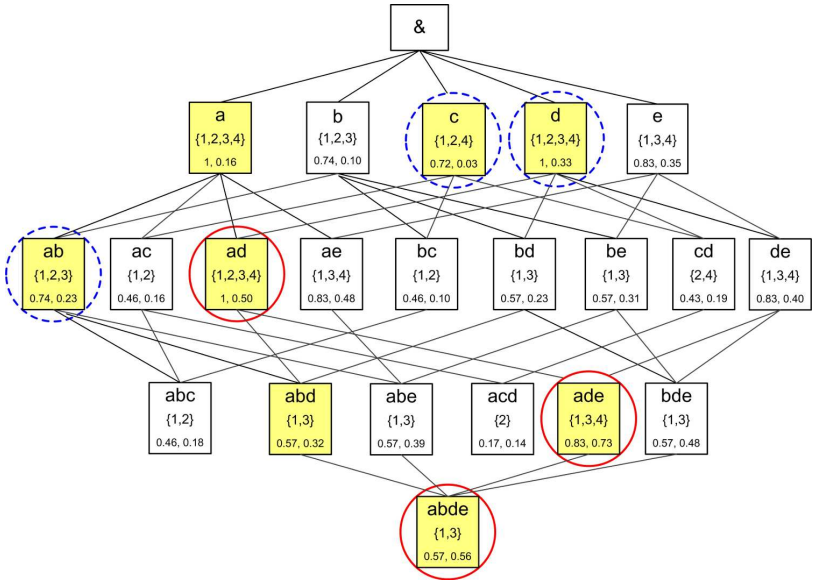


Figure 3: A clique lattice with the graphs from Fig. 2. Each rectangle contains a clique (e.g., "ab"), a corresponding set of graphs that the clique belongs to (e.g., $\{1, 2, 3\}$), a graph-weight-support (e.g., $support^g(C) = 0.74$ via Eq.(3)) and a weight-support (e.g., $support^w(C) = 0.23$ via Eq.(2)). The rectangles in yellow denote the DFS 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 HGWCC with $\theta^g = 0.5$, among which the rectangles with solid red circles are the final HWCC with $\theta^w = 0.5$.

Two types of distance are considered in our correlation estimation,

(1) 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.

(2) The great-circle distance [9], $d_{ij}^{gc} = r\,\Delta\hat{\sigma}_{ij}$, 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 that i and j belong to,*

$$c_{ij}^w = \frac{\sum_{C\in\mathcal{C}} I((i,j) \subseteq C)\, support^w(C)}{\max_{ij} \sum_{C\in\mathcal{C}} I((i,j) \in C)\, support^w(C)}, \qquad (5)$$

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

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

2.3 The Prediction Algorithm

We take the following pseudo codes 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 Eq.(4) or Eq.(5) using the mined HWCC; and (3) Run kNN or LapRLS algorithm to find the top k likely outbreak habitats.

The two machine learning methods *kNN* and LapRLS are explained as follows. We hypothesize that *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. 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 and obtain the top k habitats with the largest correlation based on the k nearest neighbor method (kNN). For example, if "a" in Fig. 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

Laplacian based regularized least-square method (LapRLS) [1], where the normalized Laplacian matrix $\mathcal{L}$ is calculated based on a habitat correlation matrix $\mathbf{W} = [c_{ij}^w] \in \mathbb{R}^{n \times n}$,

$$\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}, \text{ where } \mathbf{D} = \text{diag}(\mathbf{W1}),$$

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,

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

where $\boldsymbol{f} \in \mathbb{R}^{n \times 1}$ is the prediction vector, $\boldsymbol{y}$ is the label vector with $y_i = \begin{cases} 1 & \text{if } i = p, \\ 0 & \text{if } i \neq p. \end{cases}$. Above, $||\cdot||_F$ denotes the Frobenius norm and α is the tradeoff parameter. So, 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, since it can propagate the label via local connections [1], which is also supported by our experimental results in Section 3.2.

3 Experiments

3.1 Data Collection

Our on-site studies were conducted at the Qinghai Lake National Nature Reserve, Qinghai Province, China, between March 2007 and December 2009. 59 birds were selected randomly from different flocks to tie a battery powered GPS device to each of them. More details of the data are presented in Table 1. We had collected nearly one million migration records by December 25, 2009. We selected those 29 bar-headed geese in our subsequent analysis for the same type of birds. Finally, we have 103 habitats (i.e., nodes in graphs) and 29 graphs (one for each bird).

The reverse transcription-polymerase chain reaction (RT-PCR) [1] technique was used to confirm whether a bird is or not infected with the virus. All of the samples were immediately placed into small tubes containing transferring solution and then stored in a container of liquid nitrogen within two hours. We tested 1,055 samples by using RT-PCR, and 12 bar-headed geese, three ruddy shelducks and 14 brown-headed gulls were confirmed to be positive for an H5N1 subtype. These data are shown in the last column of Table 1, which indicates that the prevalence of H5N1 in Qinghai Lake was high. In order to obtain the relationship between migratory birds and H5N1 outbreaks, information about H5N1 outbreaks were extracted from the Ministry of Agriculture of the People's Republic of China Database and OIE Database for the period of February 2004 - May 2009.

3.2 Summary of Experimental Results

3.2.1 H5N1 Outbreak Analysis Using Mined Cliques

In this section, we focus on result analysis of clique mining. We applied the HELEN algorithm to those 29 graphs and 103 vertices to extract cliques. One high-weight clique C_{15} is shown in Fig. 4. If we only consider its frequency support ($support^f = 3/29$), C_{15} would be pruned. However, the clique has a weight of 0.13, 0.16 and 0.052, respectively, according to $W_{frequency}$, W_{time} and $W_{density}$ weighting strategies, and contributes to more than 5.2% of the total time of the birds' spring migration time. The table in Fig. 4 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 outbreak. The weight of those habitats does reflect the possibility of virus transmission.

From the above analysis, we can see that high-weight closed-clique mining can help biological professionals make better decisions, e.g., highlight some high-weighted cliques. More importantly, we discovered that 24% of mined cliques have low frequency but high weighted support, which shows the importance of weight clique mining, since otherwise, these low frequency cliques would be pruned by the traditional frequent closed clique mining algorithms. More mining results can be found at Link[1].

3.2.2 H5N1 Outbreak Prediction Using Mined Cliques

In this section, we describe our prediction experiments. We reserve all of the 245 cliques that are mined from those 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 $support^f$, respectively. Among those 103 habitats, 16 habitats have been reported one or more cases

[1]World Health Organization. Recommendations and laboratory procedures for detection of avian influenza A(H5N1) virus in specimens from suspected human cases.
http://www.who.int/csr/disease/avian_influenza/guidelines/labtests/en/index.html

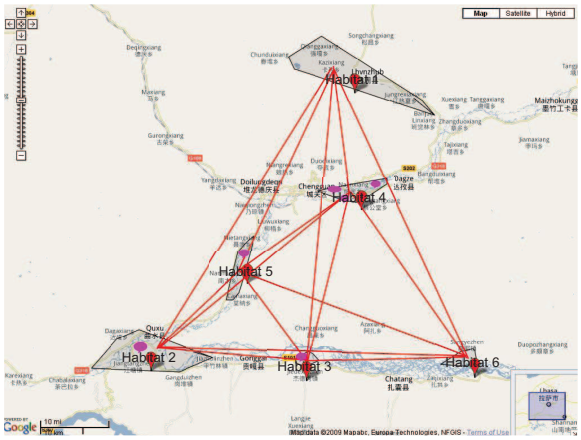
[1]Link: www.qinghailake.csdb.cn/qhlakesdm/page/paper/link1.htm
[2]Link: www.qinghailake.csdb.cn/qhlakesdm/page/paper/link2.htm
[3]Link: www.qinghailake.csdb.cn/qhlakesdm/page/paper/link3.htm
[4]Link: www.qinghailake.csdb.cn/qhlakesdm/page/paper/link4.htm
[5]Link: www.qinghailake.csdb.cn/qhlakesdm/page/paper/link5.htm

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

Bird type	bird number	Active time		Stay (days)		Migration record number	H5N1 rate (RT-PCR)
		Start	End	Max	Min		
bar-headed geese	29	2007-03-21	2009-10-21	745	48	783,240	2.27% (12/528)
ruddy shelduck	20	2007-03-21	2009-02-01	347	28	179,302	2.17% (3/138)
brown-headed gull	10	2007-06-21	2008-06-07	159	41	37,242	3.60% (14/389)

Table 2: The H5N1 outbreak prediction performance of HELEN-p(kNN) using habitat correlation estimated from geometric locations and migration data of bird satellite tracking system. Note, Pre@k=$\frac{\#\textit{positive habitat}}{k}$, threshold $\theta^w = 0$ and $\alpha = 1$.

	Geometric locations		*Using bird satellite tracking system*							
	HELEN-p(kNN)		HELEN-p(kNN), c_{ij}^w				HELEN-p(LapRLS), c_{ij}^w			
	c_{ij}^{gc}	c_{ij}^{ec}	$W_{freq.}$	W_{time}	$W_{dens.}$	$support^f$	$W_{freq.}$	W_{time}	$W_{dens.}$	$support^f$
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$_{\pm 0.23}$	**0.85**$_{\pm 0.27}$	0.76$_{\pm 0.08}$	0.84$_{\pm 0.13}$	**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$_{\pm 0.13}$	**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$_{\pm 0.08}$	**0.50**$_{\pm 0.04}$	0.42$_{\pm 0.03}$	0.42$_{\pm 0.03}$	0.48$_{\pm 0.04}$



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

Figure 4: A mined high-weight closed clique, C_{15}, with low frequency support ($support^f = 3/29$). Detailed information of the habitats and weight about the clique C_{15} are shown in the table.

of H5N1 outbreaks, i.e., they are *positive habitats*. In each prediction test, we take one *positive habitat* out of those 16 habitats, and report the averaged results over 16 times.

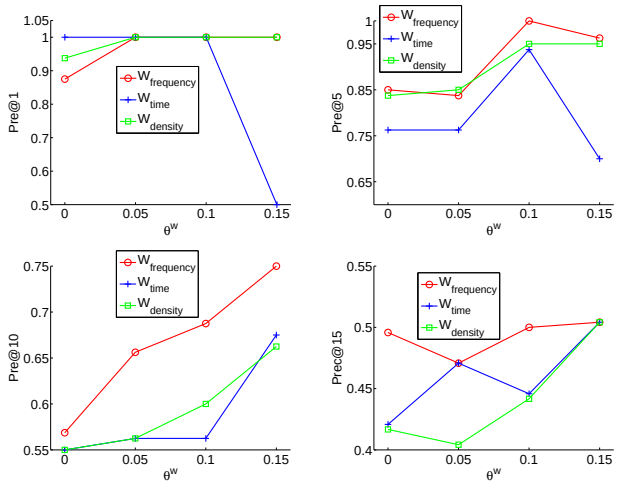


Figure 5: The H5N1 outbreak prediction performance of HELEN-p(LapRLS) with different values of θ^w.

To gain more insights on HWCC and the effect of the support threshold θ^w, we first study the prediction performance with $\theta^w = 0$, and then increase its value gradually with $\theta^w \in \{0.05, 0.1, 0.15\}$.

The prediction results with $\theta^w = 0$ are shown in Table 2, from which we can have the following observations: (1) the approach of using clique-based correlation is much better than that using the habitats' geometric information, which clearly shows the usefulness of the bird satellite tracking system or migration network in habitat correlation estima-

tion; and (2) although the clique-based correlation may fail to build connections of 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 Link[3], from which can see that HELEN-p(LapRLS) improves the prediction performance and beats kNN in all cases.

The prediction performance of HELEN-p(LapRLS) with different values of θ^w are shown in Fig. 5. We can see that, (1) using a relatively larger threshold further improves the prediction performance in most cases, and this effect can be explained by the fact that a reduction of noise in the clique weights results in a better correlation estimation in Eq.(5); and (2) using a too large threshold may reduce the prediction performance, which makes sense since the correlation between two habitats may not appear when using too few selected closed cliques. We can conclude that using a relatively higher threshold is better in prediction, which supports our assumption that H5N1 spreads via high-weight closed cliques.

4 Conclusions and Future Work

In this paper, we have developed a novel H5N1 outbreak prediction algorithm (HELEN-p). In particular, we make use of the mined cliques and machine learning methods for H5N1 outbreak prediction. The experimental results show that the mined cliques, habitat correlation calculations and machine learning methods can greatly assist biologists in H5N1 outbreak analysis and prediction. More importantly, our assumption that *H5N1 spreads via high-weight closed cliques and frequent cliques* is also supported by the experimental results (see Link[1] and Link[2] for more results). For future work, we are interested in exploiting some sophisticated algorithms to integrate different weighting strategies, where some preliminary results using linear combinations are shown at Link[4] and Link[5].

Acknowledgement

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No. 61003138 and 91224006, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA06010202 and XDA05150401, and the Hong Kong RGC Project under Grant No. 621010.

References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

[2] H. Chen, G. J. D. Smith, S. Y. Zhang, K. Qin, J. Wang, K. S. Li, R. G. Webster, J. S. M. Peiris, and Y. Guan. Avian flu: H5n1 virus outbreak in migratory waterfowl. *Nature*, 436:191–192, July 2005.

[3] Yuan-Sheng Hou, Yu-Bang He, Zhi Xing, Peng Cui, Zuo-Hua Yin, and Fu-Min Lei. Distribution and diversity of waterfowl population in qinghai lake national nature reserve. *ACTA ZOOTAXONOMICA SINICA*, 34(1):184–187, 2009.

[4] Juthatip Keawcharoen, Debby van Riel, Geert van Amerongen, Theo Bestebroer, Walter E. Beyer, Rob van Lavieren, Albert D.M.E. Osterhaus, Ron A.M. Fouchier, and Thijs Kuiken. Wild ducks as long-distance vectors of highly pathogenic avian influenza virus (h5n1). *Emerging Infectious Diseases*, 14(4):600–607, April 2008.

[5] Zheng Kou, Yongdong Li, Zuohua Yin, Shan Guo, Mingli Wang, Xuebin Gao, Peng Li, Lijun Tang, Ping Jiang, Ze Luo, Zhi Xin, Changqing Ding, Yubang He, Zuyi Ren, Peng Cui, Hongfeng Zhao, Zhong Zhang, Shuang Tang, Baoping Yan, Fumin Lei, and Tianxian Li. The survey of h5n1 flu virus in wild birds in 14 provinces of china from 2004 to 2007. *PLoS ONE*, 4(9):e6926, 2009.

[6] Jinhua Liu, Haixia Xiao, Fumin Lei, Qingyu Zhu, Kun Qin, Xiaowei Zhang, Xinglin Zhang, Deming Zhao, Guihua Wang, Youjun Feng, Juncai Ma, Wenjun Liu, Jian Wang, and George F. Gao. Highly pathogenic h5n1 influenza virus infection in migratory birds. *Science*, 309(5738):1206, August 2005.

[7] MingJie Tang, Weihang Wang, Yexi Jiang, Yuanchun Zhou, Jinyan Li, Peng Cui, Ying Liu, and Baoping Yan. Birds bring flues? mining frequent and high weighted cliques from birds migration networks. In *DASFAA (2)*, pages 359–369, 2010.

[8] Mingjie Tang, Yuanchun Zhou, Jinyan Li, Weihang Wang, Peng Cui, Yuanseng Hou, Ze Luo, Jianhui Li, Fuming Lei, and Baoping Yan. Exploring the wild birds' migration data for the disease spread study of h5n1: a clustering and association approach. *Knowl. Inf. Syst.*, 27:227–251, May 2011.

[9] Thaddeus Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23(176):88–93, 1975.

Yuanchun Zhou is an associate professor and director assistant of Scientific Data Center at the Computer Network Information Center, Chinese Academy of Sciences. He got his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. His main research interests include bid data mining, cloud computing, data intensive computing and applications. He has over 60 scientific publications in international conference and journal.

Mingjie Tang was born in China. He received his BS degree in Computer Science from Sichuan University in 2007, and master Degree from Graduate University Chinese Academy of Sciences (CAS) at Beijing in 2010. Currently, He is a graduate student at Purdue University. His recent work focuses on database, big data analysis and data mining.

Weike Pan received the Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology in 2012. He is currently a post-doctoral research fellow at the Department of Computer Science, Hong Kong Baptist University. He is also the information officer of ACM Transactions on Intelligent Systems and Technology (TIST). He was a senior engineer at Baidu Inc.

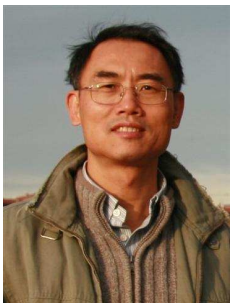
Jinyan Li is an Associate Professor and core member at Advanced Analytics Institute and Center for Health Technologies, Faculty of Engineering and IT, University of Technology, Sydney, Australia. His research is focused on fundamental data mining algorithms, machine learning, gene expression data analysis, structural bioinformatics, and information theory. He is known for the notion of emerging patterns in data mining, and is known for ‘double water exclusion hypothesis in bioinformatics. Jinyan obtained his PhD from the University of Melbourne, Master degree of Engineering from Hebei University of Technology, and Bachelor degree of Science from National University of Defense Technology.

Weihang Wang is a Ph.D. student at Purdue University. Before she came to Purdue, she won her master degree in computer science from Graduate University of Chinese Academy of Sciences in 2010, and won her bachelor degree in computer science from China Agricultural University in 2007. Now she is doing research on distributed system, cloud computing and network.

Jing Shao received his BS degree in Computer Science from Guizhou University in 2009, and master Degree from Graduate University Chinese Academy of Sciences (CAS) at Beijing in 2012. His research interests include moving object mining, big data mining and data analyzing.

Liang Wu is a master student at the Computer Network information center, Chinese Academy of Sciences. His research interests include data mining and machine learning.

Jianhui Li is a professor at the Computer Network Information Center, Chinese Academy of Sciences. He got his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2007. His main research interests include big data mining, large-scale distributed databases management and integration, semantic-based data integration, data intensive computing and scientific application.



Prof. Qiang Yang is the head of Huawei Noahs Ark Lab in Hong Kong. He has been a professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology (HKUST) since 2007. Prior to joining HKUST, he had been a faculty member at the University of Waterloo and Simon Fraser University in Canada. He is an IEEE Fellow, IAPR Fellow and ACM Distinguished Scientist. His research interests are data mining and artificial intelligence. Qiang received his PhD from the University of Maryland, College Park in 1989. His research teams won the 2004 and 2005 ACM KDDCUP competitions on data mining. He is the vice chair of ACM SIGART, the founding Editor in Chief of the ACM Transactions on Intelligent Systems and Technology (ACM TIST), and organizer for many international conferences and workshops, including the PC Co-chair for ACM KDD 2010, the General Chair for ACM KDD 2012 in Beijing and PC Chair for IJCAI 2015 Conference in Argentina.

Baoping Yan is a professor and chief engineer in the Computer Network Information Center, Chinese Academy of Sciences. Prior to this, she served as vice president of Dawning Computer Co., a famous computer high-tech company in China. She has completed analysis and design of computer network system, research and implementation of industrial automation and CIMS network technology, ATMbased workstation cluster system, standard management of large-scale network and system integration, Internet/Intranet comprehensive information management system, etc. Currently, she is responsible for the planning and construction of the informatization of the Chinese Academy of Sciences during the 10th Five-year Plan period. She has published over 100 research papers at home and abroad. The government has granted her special allowance for her outstanding contributions.