OPEN

Emerging Microbes and Infections (2013) 2, e89; doi:10.1038/emi.2013.91

www.nature.com/emi

ORIGINAL ARTICLE

Silent geographical spread of the H7N9 virus by online knowledge analysis of the live bird trade with a distributed focused crawler

Chen Chen[1,2,*], Shan Lu[1,2,*], Pengcheng Du[1,2,*], Haiyin Wang[1], Weiwen Yu[1], Huawen Song[1] and Jianguo Xu[1,2]

Unlike those infected by H5N1, birds infected by the newly discovered H7N9 virus have no observable clinical symptoms. Public health workers in China do not know where the public health threat lies. In this study, we used a distributed focused crawler to analyze online knowledge of the live bird trade in first-wave provinces, namely, Jiangsu, Zhejiang, Anhui, and Shanghai, to track the new H7N9 virus and predict its spread. Of the 18 provinces proposed to be at high risk of infection, 10 reported human infections and one had poultry specimens that tested positive. Five provinces (Xinjiang, Yunnan, Guizhou, Shaanxi, and Tibet) as well as Hong Kong, Macao, and Taiwan were proposed to have no risk of H7N9 virus infection from the live bird trade. These data can help health authorities and the public to respond rapidly to reduce damage related to the spread of the virus.
Emerging Microbes and Infections (2013) **2,** e89; doi:10.1038/emi.2013.91; published online 18 December 2013

Keywords: H7N9; online knowledge; outbreak

INTRODUCTION

Approximately one month after confirmation of the first three cases of human infection with the new reassortant avian influenza virus (H7N9), the disease had extended to 42 cities in 12 provinces, infecting 134 people and killing 45 people in China as of September 10, 2013.[1] The viral isolates from patients were similar to the isolate from an epidemiologically linked market chicken.[2] Approximately 77% of patients had a history of exposure to live poultry.[3] The virus was clearly of avian origin, and the patients were infected in farmers' markets with live birds (FMLB) through an unknown mechanism. Suspending the FMLB can prevent new human infections in cities with a high population density.[4] The sporadic human infection patterns were distributed in 39 cities and occurred at various times, indicating that the virus is silently spreading in live birds over a far larger geographical area and at a far greater speed than originally thought.[2,5,6] The movement of live birds is a well-known risk factor for the geographic dissemination of the virus among poultry flocks. The daily incidence of H5N1 virus outbreaks in Vietnam peaks around the annual holiday festivities in February coinciding with poultry movement increases.[7,8] Therefore, determining the location of live birds carrying the virus and the FMLB that is contaminated by the live-bird trade is critical in stopping the surge of human infection and the geographical dissemination of the disease. The traditional investigation of animals carrying the virus is based on the virological diagnosis of animals with a possible epidemiological link.[9,10] This task can be difficult because China has approximately six billion domestic birds. Until the source of infection has been identified and controlled, more cases of human infection by the virus are expected.[6,11] Crowd-powered expansion and the distributed focused crawler are methods that use large-scale online data to reveal human behavior accessible to research.[12] The method as used in this study to reveal live-bird-trading information to predict the potential nationwide geographic spreading of H7N9 virus, which could not be identified by traditional epidemiological investigation methods.

MATERIALS AND METHODS

Infoepidemiology and internet privacy

Through infoepidemiology, high-throughput online data associated with the live bird trade are collected by a distributed focused crawler, and we used these data to estimate the activity of trading links between provinces or cities by analyzing the frequency of their co-occurrences in one web page. We also collected all public online information in which clear trading traces were directly observed as a control for use in further analysis. All queries used in this project can be found on a web site available to the public. The database retains no information about the identity, Internet protocol address, or specific physical location of any user.

Crowd-powered extension to obtain information on trade between cities

We used the context-based connection (CC) and information-based connection (IC) to evaluate the activity of live poultry trade at two levels, provincial and city, respectively. CC data are connections directly observed from web knowledge collected by a crowd-powered extension system in our center. The key words and 367 city names were distributed in the discussion forum. The searchers were required to obtain all traces with the city name and live poultry trade or its synonyms. The search results were returned, collected, and combined. All

[1]State Key Laboratory for Infectious Disease Prevention and Control and National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China and [2]Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Hangzhou 310003, China
*These authors contributed equally to this work.
Correspondence: JG Xu
E-mail: xujianguo@icdc.cn
Received 11 June 2013; revised 20 September 2013; accepted 12 November 2013

queries were required to provide the source websites or internet protocol address (IP address) when they supplied connection information between cities. This information was individually validated to ensure accuracy. Our script-on-text comparison removed repeated queries and filtered ambiguous web sites. The co-occurrence of city names in one page was estimated to have trading links by comparing scripts in automatic manuscripts. All the connections between these cities were double-checked manually. A total of 315 queries and 591 CCs associated with 211 cities were included in our study.

Information-based connection by a distributed focused crawler

After CC data were collected, we first analyzed the information online using the distributed focused crawler technology from 244 public news web sites as well as 56 forums and microblogs, including common sites such as ''Sina,'' ''Sohu,'' and ''Yahoo.'' Keywords were used to select online knowledge and entered using the following format: (city name) and (live poultry or poultry or dove or quail or birds or migratory bird or chicken or duck or goose) and (come from or exchange or source or carrier or travel or supply or market or slaughterhouse or production area). The pages that fit our requirements were included in the hash list. A bloom filter was used to adjust and remove repeated information. These pages were clustered with a support vector machine model and categorized into different groups to analyze the orientation of web pages. These confirmed web pages were then downloaded and coordinated by Apache ZooKeeper (The Apache Foundation, Delaware, US).

The IC of two cities was calculated by the number of records of the web page containing both cities. Using CC data, we estimated the error rate of IC data under the assumption that the CC data were complete. The distribution of IC with and without CC support was drawn, revealing that CC supported data with more IC records (Supplementary Figures S1A and S1B). This result suggests that our IC data were reasonable for use in the analysis of trading connections in provinces and cities. When compared to the connections between provinces, the connections between cities show sharper peaks and fewer connections because of insufficient information. Accurate analysis of trading links was difficult in these cities, especially in Western China. We defined noise as information that was collected but not relevant to construct the connection of bird trade. We used the following rules to calculate the noise rate: (i) we obtained the rough dataset by using general, relevant keywords about the infections, and then (ii) we used more specific, relevant keywords to search and generate a more accurate dataset before (iii) the noise rate was calculated as the difference between the first and second datasets. To exclude the noise information from web sites, a cutoff value was chosen and the number of connections was recorded (Supplementary Figure S1C). The false positive ratio (FPR) and false negative ratio (FNR) of IC were estimated based on CC because we lacked complete and accurate datasets for information analysis (Supplementary Figure S1D). Different cutoff curves for provinces and cities reveal cutoffs of 10 and 20 as the thresholds for cities and provinces, respectively. The correlation between each pair of provinces was calculated and compared using the number of their supported web sites (Supplementary Figure S2). Infection risk was estimated by the connections between cities and provinces using the following two parameters: (i) the number of provinces or cities potentially associated with the outbreak region and (ii) the number of queries that contain the keywords and the name of two provinces or cities.

Topological map of virus spread constructed with a force-directed graph

We designed a tool to determine all possible connections between cities using public open source code from Tim Dwyer and Thomas Jakobsen (http://bl.ocks.org/mbostock/4062045#index.html). The input data comprised a table of connection results with node columns, connected target columns, and their connections. For example, if Shanghai was the node, its connections would be ''Jiangsu'' and the number of links would be ''295'' on IC. All possible connections were constructed, recorded, and entered in a structured query language (SQL) database. The pair of connections A to B and B to A was only considered once in our analysis. A force-directed graph was constructed by the connections between two provinces and cities. We only removed all connections to provinces when we calculated the connection between cities. Thus, information for some cities with connections only to provinces was lost in this graph. When we constructed the graph of the provinces, we only considered four types of connections: (i) node province to target province, (ii) cities in the node province to the target province, (iii) node province to cities in the target province, and (iv) cities in the node province to cities in the target province. These connections were accumulated as the connections between provinces. Because the connection to diseased birds was important in our analysis, we focused exclusively on the provinces and the cities directly connected to first-wave provinces and cities. Thus, at least one component in the pair of connections (node or target) belonged to the first-wave H7N9 outbreak area. Shanghai, Jiangsu, Zhejiang, and Anhui were selected as the first-wave provinces and 11 cities in these provinces were also included. Not all live poultry carried the virus, and too much data from connections would distort our understanding of the transfer of diseased live poultry. The connections from this first-wave were not included and considered in our graph. For example, Guangdong is connected to Zhejiang, and Guangxi is connected to Zhejiang; thus, these three provinces are all included in our graph. Although we had evidence showing that Guangdong had trade information on Guangxi, we did not present the connection between Guangdong and Guangxi in our graph because this connection was not associated with the first-wave trade information. The CC and IC were used together as the connections, and IC was used as the weight for each connected line. All aforementioned filters and selections were carried out by a self-designed Java script and SQL queries before a force-directed graph was drawn. The potential transmission networks of live bird trade in cities with reported H7N9 cases were derived from the connection of cities. Cities with patients but no direct connection to first-wave cities were also connected by seeking the cities that they connected to in the figures. For example, Fuzhou was not directly connected to the first-wave cities but was connected to Nanping and Ningde. We also included the relationships between cities from the database in the figures to show their possible transmission networks.

RESULTS

Source of the data

We analyzed 835 635 web pages and 2943 pages associated with our topics. Their distributions are shown in Supplementary Figure S3 and Supplementary Table S1. Shanghai has 922 pages with live poultry trading information. Surveillance of the web sites revealed that live poultry from Shanghai came from Jiangsu, Zhejiang, and Anhui Provinces. The noise of these queries was 45.17% and was determined through the method of excluding words from our search. Thus, we collected a considerable amount of accurate data versus the amount of noise data.

First wave of H7N9 spreading

Online knowledge analysis revealed that the live birds in Shanghai were traded from Jiangsu, Zhejiang, and Anhui provinces with 84%

(571) of 682 live bird trade queries linked to Shanghai authorities stated that approximately 80% of live birds in this city were from the three aforementioned provinces.[13] Therefore, the live birds from these three provinces and Shanghai are potential sources of new infection. The date of illness onset for the first patient in Shanghai was February 18, followed by March 7, 15, and 19 for Zhejiang, Jiangsu, and Anhui provinces, respectively. Therefore, we propose that the H7N9 outbreak in China occurred in two waves separated by April 5 (Figure 1), the date when Shanghai authorities announced the suspension of FMLBs and told the public that live birds carried the newly discovered H7N9 virus and that people in FMLBs were infected. This information can significantly change the trading designations and activities for live birds that carry the virus as well as public behavior to avoid the live birds during the outbreak. In the first wave, 11 prefecture cities and 106 people were infected. In the second wave, the number of infected prefecture cities increased to 28, covering 11 county level cities, but the number of infected people was reduced to 24. The first-wave was characterized by more infected people, while the second wave featured more infected cities, indicating that the live bird trade contributed to the geographical spread of the disease.[3]

Estimating the infection risk to first-wave provinces at the province level

We drew a topological provincial map of the spread of the H7N9 virus based on live bird trade information linked with first-wave provinces. The risk of H7N9 virus infection was estimated by the live bird trade links to Shanghai, Zhejiang, Jiangsu, and Anhui Provinces (Figure 2). Only seven provinces and one municipality were linked to one of the first-wave provinces, indicating that they had a very low risk of H7N9 infection from live bird trade. They are Jilin, Gansu, Qinghai, Inner Mongolia, Hainan, Yunnan, and Heilongjiang Provinces and the municipality of Chongqing cities (Figure 2). Among these provinces, Qinghai only had context information and we did not find any connection in the crawling pages. Of the 18 provinces suggested to be at high risk, 10 had confirmed cases of infection (Figure 2). In this group, Guangdong Province had poultry specimens that tested positive for

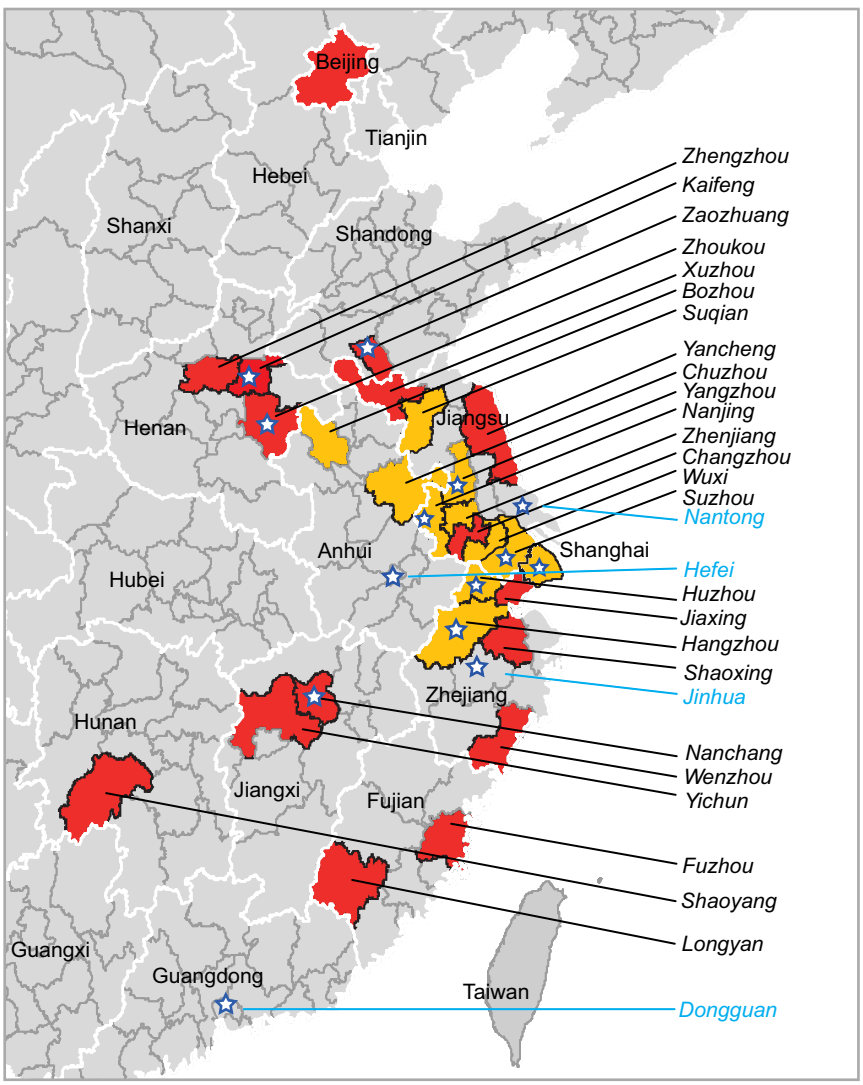

Figure 1 Geographical distribution of cities infected by the H7N9 virus in China from March to May 2013. The first day of illness onset for the first patient in each city was taken as the day when the city was infected. Infected cities were divided into two waves, separated by April 5, 2013. Cities in the first and second waves are labeled by a yellow and red circle, respectively. The cities where the virus was detected in birds are labeled by a blue star.

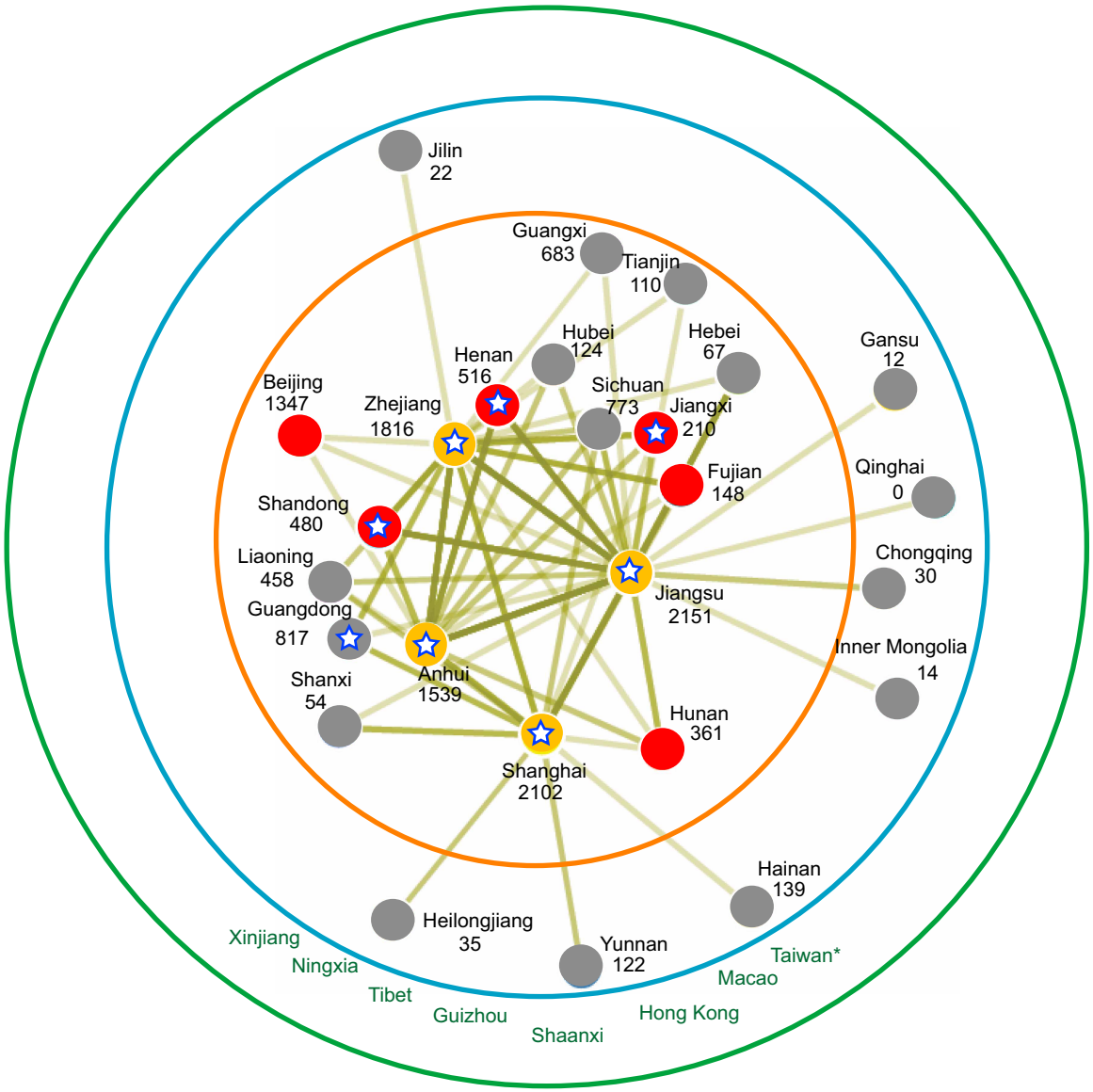

Figure 2 Provincial topological map of the spread of the H7N9 virus based on the live bird trade. A force-directed graph was drawn based on CC and IC. The infected provinces in the first (yellow node) and second (red node) waves and in provinces without infection reports (grey node) were connected according to live bird trade information weighted by the number of queries. The accumulated value to support trading information for each city is marked below its name. A concentric circle indicates the putative risk of the H7N9 virus. Provinces with birds that tested positive are labeled by a blue star.

the H7N9 virus, as reported on May 6, and a patient diagnosed on August 9. A patient diagnosed in Taiwan was infected in Jiangsu before departure. No link to first-wave provinces was observed for Xinjiang, Ningxia, Tibet, Guizhou, Shaanxi, Hong Kong, Macao, or Taiwan. Therefore, based on our analysis, no risk associated with the live bird trade in these provinces could be detected using our method.

Estimating the infection risk to first-wave provinces at the city level

We also drew a topological city map showing the spread of the H7N9 virus based on live bird trade information (Figure 3). A total of 63 cities, including 18 in which infections were reported, were clustered and connected via both knowledge-based links (i.e., IC and CC) to the first-wave cities. It should be mentioned that nine cities with diagnosed infected patients and two cities with only virus-positive birds had no knowledge-based link with the first-wave cities (Figure 3). Only four cities had live poultry specimens that tested positive for H7N9 virus, namely, Hefei (Anhui Province), Nantong (Jiangsu Province), Jinhua (Zhejiang Province), and Dongguan (Guangdong Province).

Transmission networks based on live bird trade information

The transmission networks based on live bird trade information are also shown in this study. The major links between infected cities were taken from the relationship map of cities (Figure 4A). Figure 4A shows that 8 of the 11 first-wave cities had the most frequent links to the live bird trade, indicating that these cities, namely, Shanghai, Hangzhou, Chuzhou, Yangzhou, Nanjing, Huzhou, Zhenjiang, and Wuxi, had more important functions than the others. Among them, five first-wave cities, namely, Shanghai, Hangzhou, Yangzhou, Nanjing, and Huzhou could have important functions in spreading the second wave of infection (Figure 4A). One link was found for Suqian City in Jiangsu

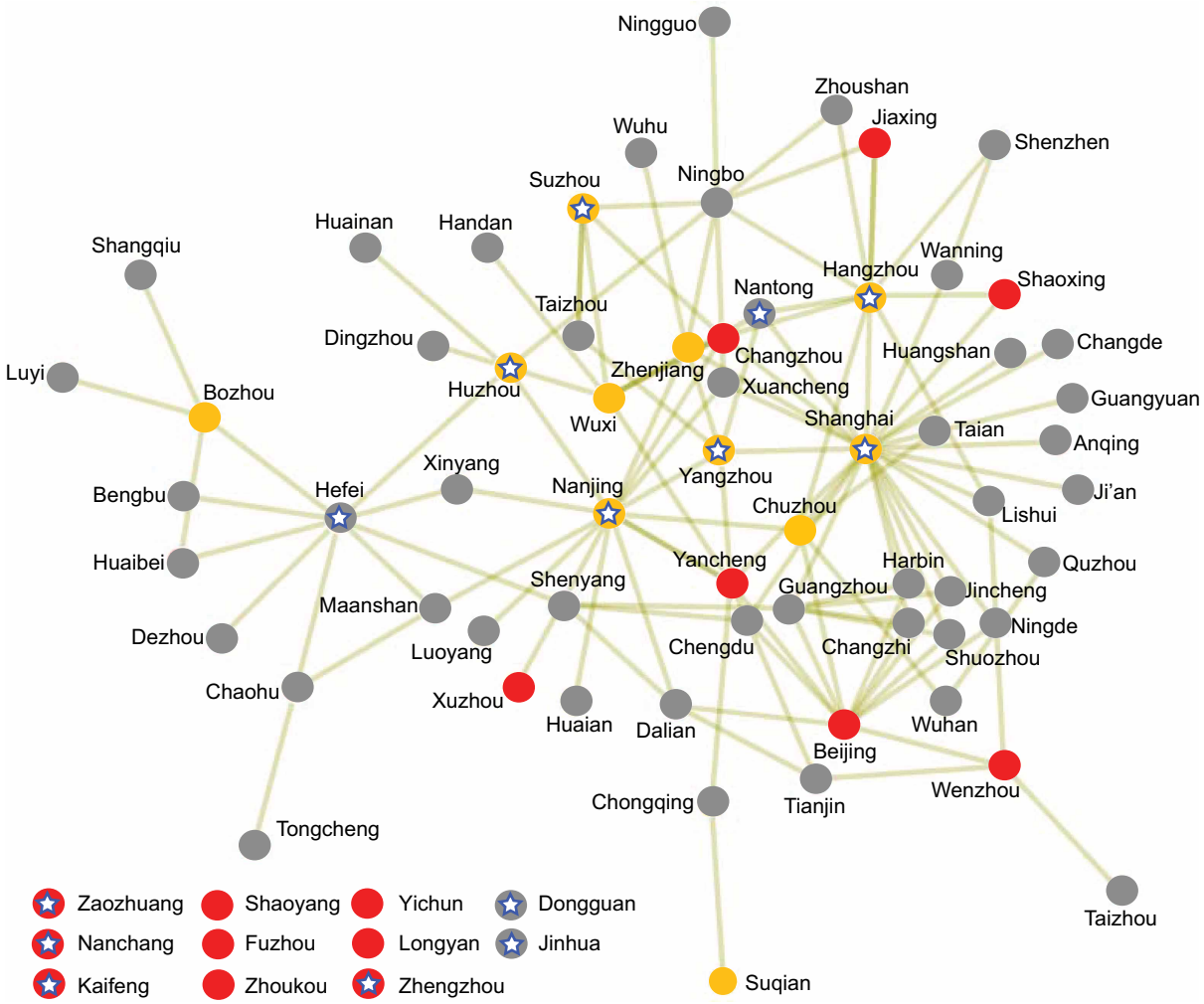

Figure 3 City topological map of the spread of the H7N9 virus based on the live bird trade. A force-directed graph was drawn based on CC and IC. Infected cities in the first (yellow node) and second (red node) waves and cities with no infection reports (grey node) were connected according to live bird trade information using information number as the weight. Cities with birds that tested positive are labeled by a blue star.

Province. Thus, the geographic spread of H7N9 virus in China was likely driven by live birds traded from the first-wave cities.

Based on online information concerning the live bird trade, we propose three transmission networks (Figures 4B, 4C, and 4D). Because the topological map has no direction, we propose a direction of transmission based on the time order, that is, when the city was infected. Figure 4B suggests that Beijing was most likely infected by live birds that carry the virus from Chuzhou (Anhui Province) or Yancheng (Jiangsu Province). Epidemiological data support the hypothesis that the two cities were infected on March 9 and April 8, respectively, and both were infected earlier than Beijing (April 11) (Figure 4B). Figure 4C shows that Chuzhou was possibly infected by live birds that carry the virus from Hangzhou (Zhejiang Province) or Shanghai. Epidemiological data also show that Chuzhou was infected later than Hangzhou and Shanghai. The genome sequences of the virus strain isolated from patients in Chuzhou were highly homologous to those isolated from a patient and a chicken from Hangzhou, as well as to one of two isolates from patients in Shanghai.[1,2] Figure 4D shows that Nanchang City (Jiangxi Province) and Fuzhou City (Fujian Province) were most likely infected by the virus from live birds traded from Shanghai. Both cities were infected later than Shanghai.

DISCUSSION

Information on the live bird trade from infected cities and provinces is critical to prevent and control the H7N9 virus outbreak in China. However, this information is not available and cannot be obtained by classical epidemiology methods with limited information. The information generated from virological investigation of live birds and FMLB is insufficient to help control the spread of the virus. Suspending FMLBs can prevent new human infections in cities with a high population density. However, this intervention is insufficient to control the disease. With the number of infected cities increasing to 39, each city reported only a few infected patients, which is alarming. We must also reduce the number of infected cities to limit the geographical spread of the disease. This goal can only be achieved by stopping the geographical movement of live birds that carry the virus.

In this study, we used online knowledge of the live bird trade in conjunction with distributed focused crawler technology to study the infectious sources, transmission networks, and geographical spread of the newly discovered virus, which has national and international public health significance. This study provides valuable information for the control and prevention of the current H7N9 virus outbreak, and serves as a basis for future outbreak investigations.

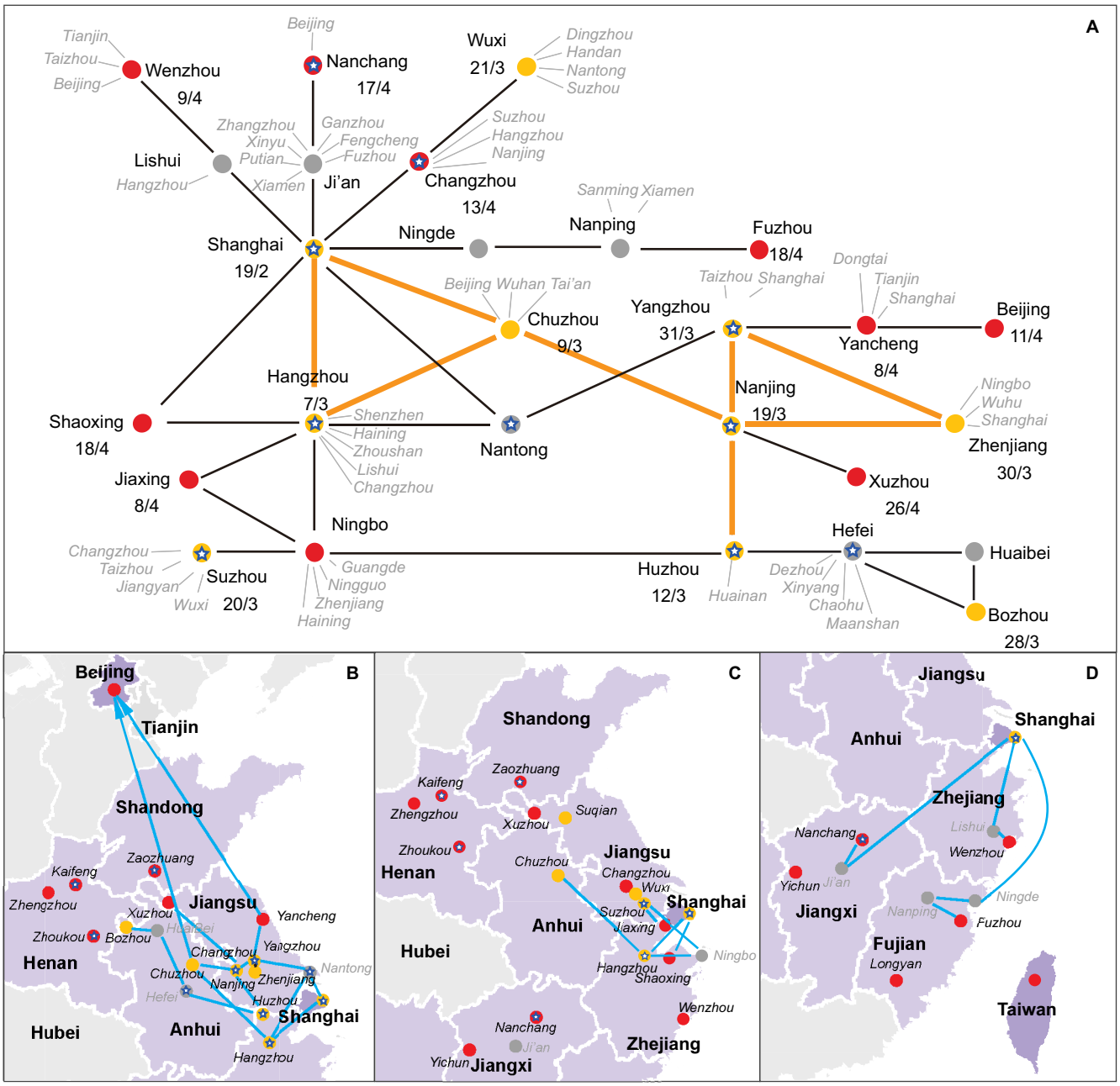

Figure 4 Proposed transmission networks of the H7N9 virus based on live bird trade information. Cities in different infection waves are represented by red, yellow, and grey nodes. Cities where the virus was detected in birds are labeled by a blue star. (**A**) Topological connection structure of infected cities based on live bird trade information. Trading among first-wave cities is indicated by the yellow line. The other cities with trade information to the cities in these networks are complemented beside the cities, and connected by the gray line. The day when the city was infected was marked as day/month below the name of the city. (**B**) Proposed transmission network for Beijing. (**C**) Proposed transmission network for Chuzhou, Anhui Province. (**D**) Proposed transmission network for Fuzhou City, Fujian Province.

This method can also be used as a measure of other public interests or concerns about health-related events. This type of information is usually difficult or impossible to obtain by traditional investigation methodology, such as face-to-face or telephone inquiry, especially for infectious diseases or other events with potential significant economic impact. Parts of the virus transmission network that we derived are supported by preliminary epidemiological observations and genome sequence analysis results from virus strains isolated from both patients and poultry. With increasing genome sequences and epidemiological information being published, new technology for infectious disease outbreak investigations will greatly improve. The limitation of this analytical method used here is that we do not know what fraction of total poultry movements could be traced through the Internet. Some information related to trading that was arranged over the phone or performed by integrated poultry companies cannot be obtained from Internet. Because viruses continue to circulate in poultry and cause no clinical signs in birds, it will be important to further improve this method using more surveillance information generated in the coming years.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Natural Science Foundation of China (81290345) and the National Key Program of Mega Infectious Diseases (2011ZX10004-001) from the Ministry of Science and Technology and National Health and Family Planning Commission, P. R. China. We thank Wentao Xia, Xiaoyong Zhao, Zhiqiang Wu, Shaohua Chen, Li Zhang and Yubiao Qiang from the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention for their generous technical support as well as software development.

1 Gao R, Cao B, Hu Y *et al*. Human infection with a novel avian-origin influenza A (H7N9) virus. *N Engl J Med* 2013; **368**: 1888–1897.
2 Chen Y, Liang W, Yang S *et al*. Human infections with the emerging avian influenza A H7N9 virus from wet market poultry: clinical analysis and characterisation of viral genome. *Lancet* 2013; **381**: 1916–1925.
3 Li Q, Zhou L, Zhou M *et al*. Preliminary Report: Epidemiology of the Avian Influenza A (H7N9) Outbreak in China. *N Engl J Med* 2013; doi: 10.1056/NEJMoa1304617.
4 Xu J, Lu S, Wang H, Chen C. Reducing exposure to avian influenza H7N9. *Lancet* 2013; **381**: 1815–1816.
5 Horby P. H7N9 is a virus worth worrying about. *Nature* 2013; **496**: 399.
6 Parry J. H7N9 virus is more transmissible and harder to detect than H5N1, say experts. *BMJ* 2013; **346**: f2568.
7 van den Berg T. The role of the legal and illegal trade of live birds and avian products in the spread of avian influenza. *Rev Sci Tech* 2009; **28**: 93–111.
8 Soares Magalhaes RJ, Ortiz-Pelaez A, Thi KL *et al*. Associations between attributes of live poultry trade and HPAI H5N1 outbreaks: a descriptive and network analysis study in northern Vietnam. *BMC Vet Res* 2010; **6**: 10.
9 Kan B, Wang M, Jing H *et al*. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J Virol* 2005; **79**: 11892–11900.
10 Wang M, Di B, Zhou DH *et al*. Food markets with live birds as source of avian influenza. *Emerg Infect Dis* 2006; **12**: 1773–1775.
11 Alcorn T. As H7N9 spreads in China, experts watch and wait. *Lancet* 2013; **381**: 1347.
12 Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. *Sci Rep* 2013; **3**: 1684.
13 Shanghai Municipal Agricultural Committee. *Poultry and eggs produced locally are safe*. Shanghai: SHAC, 2013. Available at http://www.shagri.gov.cn/xwkd/news/201304/t20130409_1340203.htm (accessed 1 May 2013).

Supplementary Information for this article can be found on *Emerging Microbes & Infections*' website (http://www.nature.com/EMI/)