



Gaining Insights into Epidemics by Mining Historical Newspapers

**E. Thomas Ewing, Samah Gad,
and Naren Ramakrishnan**
Virginia Tech

A collaborative project combines innovative algorithmic techniques with traditional textual analysis of newspapers to learn more about past health events, guiding preparedness for future crises.

Health authorities are continually on the alert for reports of deadly disease outbreaks, such as the recent H7N9 avian flu epidemic in China, to prepare for a possible pandemic and coordinate an effective response at local, regional, and global levels.

Analyzing historical accounts of earlier pandemics can provide useful insights into disease transmission methods, community vulnerability, the efficacy of different public health interventions, and other questions. Toward this end, a recently initiated project is applying novel data mining techniques to a corpus of digitized newspapers published in 1918 during the Spanish flu pandemic to better understand the flow of information about the impact of that deadly disease.

PROJECT GOALS

An Epidemiology of Information (www.flu1918.lib.vt.edu) brings together faculty and graduate students in the humanities, rhetoric, computer science, history, public health, and information sciences at Virginia Tech. Combining algorithmic techniques with interpretive analytics, the project makes use of more than 100 newspapers for 1918 available in two repositories: *Chronicling America* at the US Library of Congress, and *Peel's Prairie Provinces* at the University of Alberta.

Here we focus on the former collection, which includes digitized newspapers from across the US and is freely available for public use. The API for *Chronicling America* enables the easy extraction of text captured through optical character recognition (OCR) technology and

is thus amenable to data mining methods.

During the 1918 Spanish flu pandemic, newspapers were the only widely circulated media. As Figure 1 shows, they recognized both the severity of the epidemic and the importance of timely public health responses. Newspapers published accounts of the disease's impact on local communities and used the existing network of wire services to report on other locations.

To better understand the influence of newspaper reports on efforts to control the spread of the pandemic, we integrate data mining and network analysis methods with traditional textual analysis. Specifically, we're exploring the hypothesis that extensive newspaper reporting paved the way for public acceptance of stringent, and only partially effective, countermeasures.

DYNAMIC TEMPORAL SEGMENTATION

We first used a dynamic temporal segmentation algorithm to map the changing geography of newspaper reports on Spanish flu, focusing on articles from the *New York Herald Tribune*. The algorithm aims to find time boundaries between maximally different topics, signifying qualitative shifts in newspaper coverage. It models topics using latent Dirichlet allocation (LDA) and discovers the segmentation boundaries by optimizing for the Kullback-Leibler divergence between topic distributions.

Figure 2 displays clusters of topics in newspaper reports on Spanish flu from 12 September to 1 November 1918 discovered by our algorithm and segmented into two time periods. Visual analysis of these clusters indicates the prominence of disease terms such as “germ,” “health,” “epidemic,” “pneumonia,” “death,” and “hospital.” However, it also illustrates the importance of locations, which change over time. For example, “Massachusetts,” the location of Camp Devens, and “Dix,” an army camp in New Jersey, appear in the first but not the second segment. By calling attention to the relative

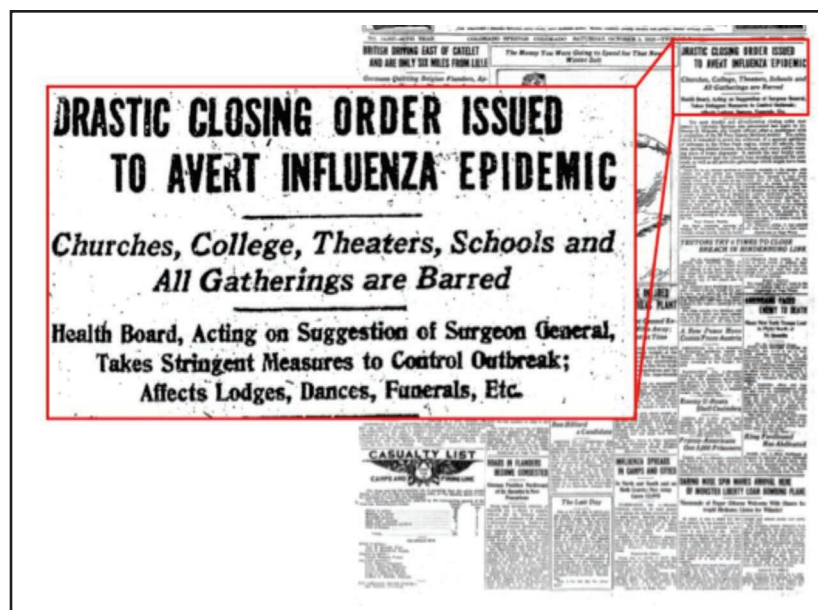


Figure 1. A front-page article in the 5 October 1918 edition of the *Colorado Springs Gazette* announces local public health measures implemented in response to the Spanish flu pandemic.

significance of geographical terms in different time periods, these visualizations mark the disease’s spread to new sites.

NETWORK ANALYSIS

To trace the relationship between the spread of Spanish flu and the spread of information about the disease, we studied more than 700 newspaper reports from 52 papers

published between 15 September and 25 October about influenza outbreaks at six different locations: Camp Devens, Camp Dix, Camp Lee (in Virginia), Philadelphia, New Orleans, and Seattle.

We found 194 articles alone about Camp Devens, where more than 17,000 influenza cases resulted in 787 deaths. The number of articles peaked during a five-day period

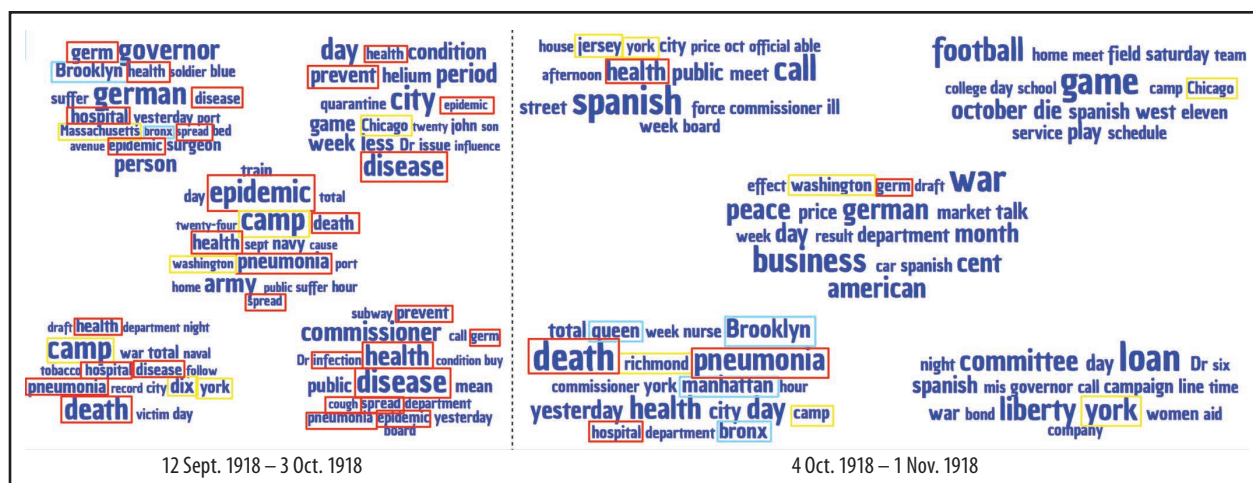


Figure 2. Clusters of topics discovered in newspaper reports on Spanish flu, segmented into two time periods. Red boxes indicate disease terms, blue boxes indicate New York infected sites, and yellow boxes indicate national infected sites. Such visualizations call attention to the relative significance of geographical terms in different time periods, marking the disease’s spread to new sites.

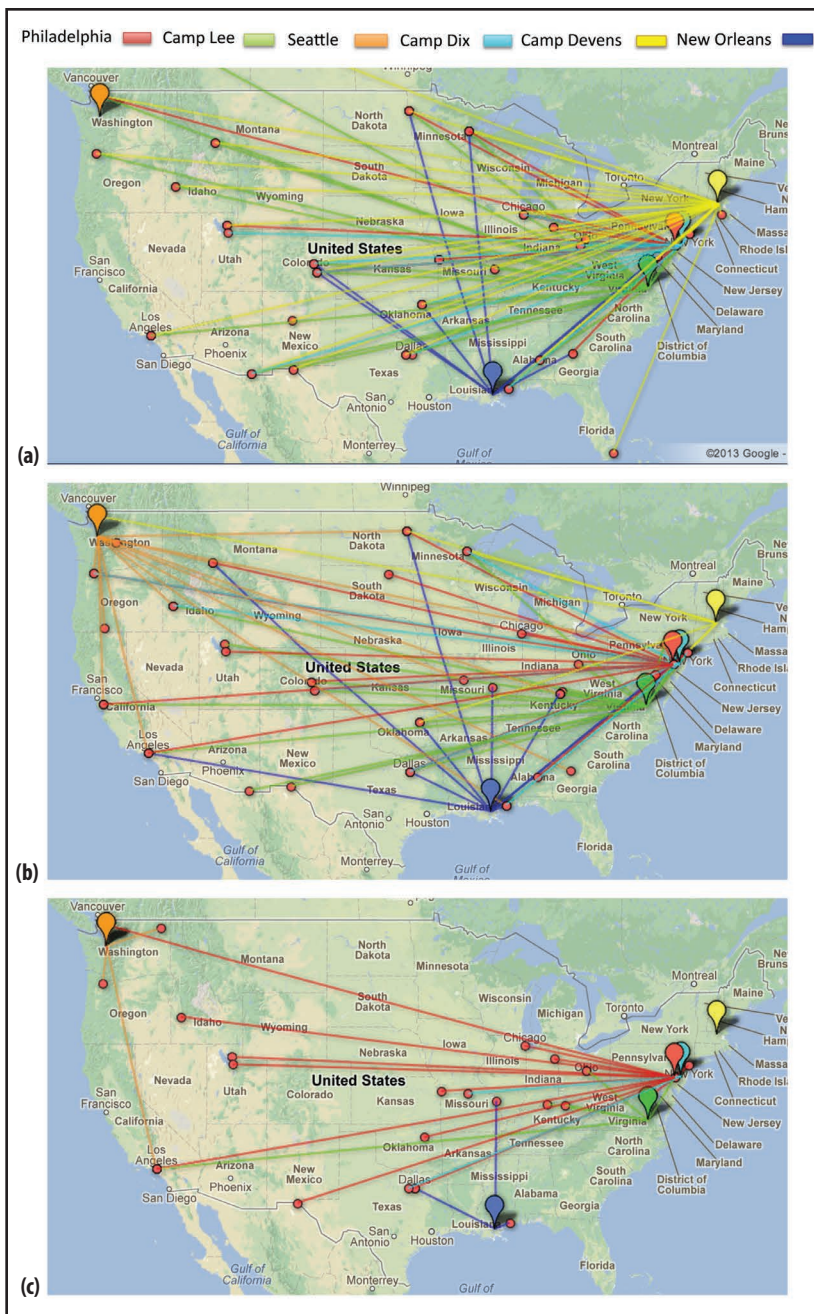


Figure 3. Mapping connections between infected and reporting sites in more than 700 reports of influenza outbreaks at six locations over time: (a) week 1 (15-21 Sept. 1918); (b) week 3 (29 Sept.-5 Oct. 1918); and (c) week 6 (19-26 Oct. 1918). The flow of information followed, but also diverged from, the course of the disease.

(20-24 September), gradually declined over the next 10 days, and fell off significantly after 5 October. This decline was due in part to a decrease in the number of cases at Camp Devens, but also to increased attention to other infected

sites, particularly those in regions covered by a particular newspaper. Major cities with early Spanish flu outbreaks, such as Philadelphia and New Orleans, also received extensive coverage nationally, whereas Seattle, where the disease did not

appear until 5 October, received primarily regional coverage.

To explain and characterize these shifts in coverage, we turned to network analysis. Network analysis methods range from graphing network connections (degree distributions, clustering coefficients) to mining network properties (communities, weak ties, bridges) to detecting key aspects of temporal and spatial evolution (compartmentalization, redistribution, coalescing of nodes). Such methods make it possible to build upon traditional humanities-based approaches that use close reading of selected texts to interpret historical experiences and identify meaningful patterns.

We mapped the connections between infected and reporting sites to variations in clustering coefficients and associated local graph properties. As Figure 3 shows, the flow of information followed, but also diverged from, the course of the disease over six weeks.

ENHANCING KNOWLEDGE DISCOVERY

We combined these network analysis results with textual analysis to test the hypothesis that extensive newspaper coverage of the pandemic enabled local health authorities to pave the way for rigorous interventions such as closing schools, churches, and theaters, banning public assemblies, restricting shopping hours, or requiring masks to be worn in public places. We illustrate our approach by looking closely at reports in one regional newspaper, the *Colorado Springs Gazette*.

From the second week of September to the end of October 1918, this newspaper published more than 200 articles about the Spanish flu pandemic, including 10 about Camp Devens. Over an eight-week period in September and October, there were more than 400 references to infected sites. As Figure 4

shows, during the first two weeks, all of the reports were about infected sites in the US outside of Colorado. By the last week of September, however, the newspaper included articles about infected sites in the state—primarily universities where soldiers attending training camps were becoming ill.

The tipping point in coverage occurred during the second full week of October, when the number of reports on infected Colorado sites exceeded the number of reports on infected sites outside the state. By the end of October, the number of reports on Colorado sites was more than double the number of reports on national sites. In other words, as the disease came closer to Colorado Springs, connections tightened to infected sites in the state and loosened to infected sites outside the state.

Textual analysis of these newspaper accounts confirms an important shift in reporting as the disease approached the reporting site.

On 1 October, before any influenza cases had been confirmed in Colorado Springs, a city health official cautioned against “relaxed vigilance on the part of every man, woman and child,” and pledged that his office had taken “every possible precaution” to prevent the spread of the disease. “If there is cooperation among all citizens,” he asserted, “there will be no epidemic.” Just two days later, however, the newspaper reported on “the epidemic which is now sweeping Colorado Springs,” even though only three patients had fallen ill “and in each instance the patients are much improved.”

On 5 October, a front-page headline in the *Colorado Springs Gazette* announced a “drastic closing order” applying to churches, theaters, and schools and a ban on all public assemblies, including dances and funerals, to control the outbreak (see Figure 1). City health officials implemented this significant intervention even though the actual

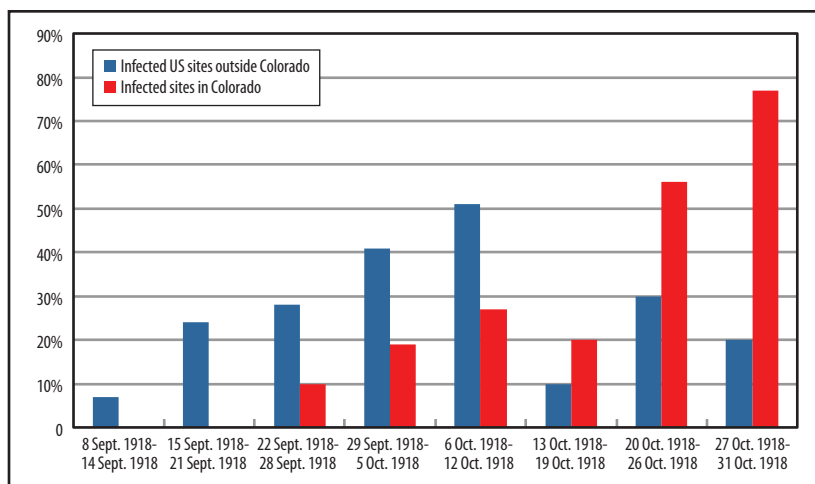


Figure 4. Distribution of reports about Spanish flu in the *Colorado Springs Gazette* from the second week of September 1918 to the end of October 1918. As the disease came closer to the city, connections tightened to infected sites in Colorado and loosened to infected sites in the US outside the state.

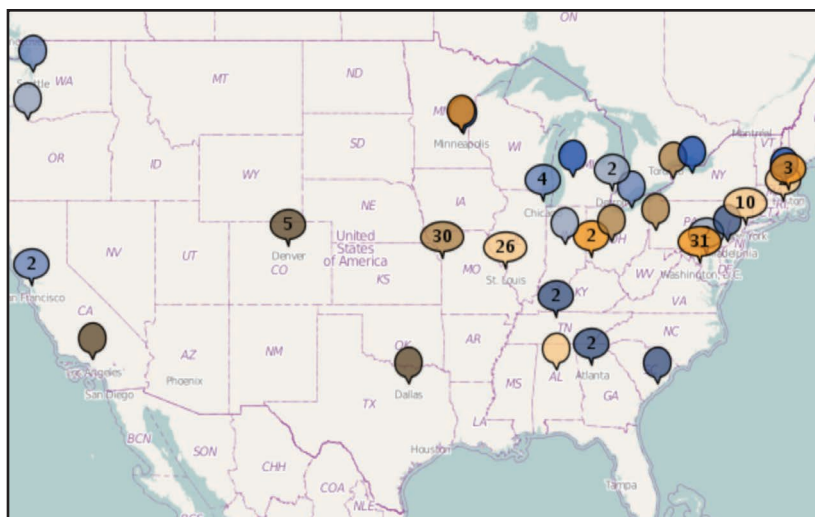


Figure 5. Distribution of 141 reports on major urban sites with widespread Spanish flu cases in *The Evening Missourian*, September-October 1918. One-third of the articles concerned St. Louis and Kansas City, which were not the most afflicted cities among the 50 urban sites examined but were the ones closest in proximity to Columbia, where the newspaper was published.

number of cases was relatively low because they anticipated, based on reports about Spanish flu in other cities, a major increase in infections and deaths. In this case, the flow of information preceded the actual spread of the disease, in ways that allowed—and even compelled—public health officials to take steps to mitigate its likely impact on their community.

SCALING UP ANALYSIS

To expand our analysis, we examined reports on infected sites in *The Evening Missourian*, a daily newspaper published in Columbia, the location of the University of Missouri. Using a list of 50 major cities with widespread Spanish flu cases drawn from *The American Influenza Epidemic of 1918-1919: A Digital Encyclopedia* (www.influenzaarchive.org),


we located more than 140 articles from a seven-week period beginning in the second week of September 1918.

As Figure 5 shows, these reports stretched from Boston, one of the first cities affected, to the West Coast cities of Seattle, Los Angeles, Portland, and San Francisco. Of these 141 reports, however, a total of 56—more than one-third—concerned St. Louis and Kansas City. These cities were not on the list of those most afflicted by influenza but were the ones closest in proximity to Columbia.

When *The Evening Missourian* announced the city's "fight" against influenza on 7 October, only 70 cases, all but one of them mild, had been reported. Nevertheless, city health officials shut down the University of Missouri, banned gatherings, and closed schools for a week. As in the case of Colorado Springs, these stringent countermeasures were motivated by newspaper accounts of the disease in other locations and the fear that Columbia would follow suit.

We obtained this information

using automated means, rather than keyword searching or manual analysis. Our approach thus demonstrates how humanities scholars can scale up their analysis using data mining and knowledge discovery techniques.

Innovative collaborative projects like An Epidemiology of Information can reveal new insights into major events such as the 1918 Spanish flu pandemic, guiding preparedness for future health crises. We're using the same approach to examine other content in historical newspapers—for example, the language used to describe the pandemic and its aftermath—and compare it to today's coverage. Another issue we're exploring is the tone of coverage in articles: whether they were alarmist or reassuring, and how that tone changed as the disease peaked and then lessened. 

Acknowledgments

Funding for the An Epidemiology of Information project is provided by National Endowment for Human-

ities grant HJ-50067-12 through the Digging into Data Challenge (www.diggingintodata.org).

E. Thomas Ewing is associate dean for research in the College of Liberal Arts and Human Sciences and a professor of history at Virginia Tech, as well as principal investigator and director of the An Epidemiology of Information project. Contact him at etewing@vt.edu.

Samah Gad is a doctoral candidate in computer science at Virginia Tech and a research assistant at the university's Discovery Analytics Center. Contact her at samah@vt.edu.

Naren Ramakrishnan, Discovery Analytics column editor, is the Thomas L. Phillips Professor of Engineering at Virginia Tech and director of the university's Discovery Analytics Center. Contact him at naren@cs.vt.edu.

 Selected CS articles and columns are available for free at <http://ComputingNow.computer.org>.

RSBD 2013

Rock Stars of Big Data

Register today!

computer.org/Big-Data

Tuesday, October 29, 2013

Computer History Museum, Mountain View, CA USA

Come to IEEE Computer Society's Rock Stars of Big Data to hear how executives from GE, Google, IBM, Intel, Kaiser Permanente, Netflix, and others, unleashed the power of Big Data.

Find out how to:

- Create a Big Data culture in your company
- Make Big Data projects succeed
- Use Big Data analytics to make the right decisions

- Empower employees and teams for Big Data projects
- Use analytics and artificial intelligence for effective decision-making
- Find help when internal resources aren't enough
- Understand Big Data's evolution and the human implications



IEEE  computer society