Monitoring Influenza Trends through Mining Social Media

Courtney D Corley*, Armin R Mikler*, Karan P Singh† and Diane J Cook‡
*Dept of Computer Science and Engineering, Univ of North Texas, Denton, TX, {corley|mikler}@unt.edu
†Dept of Biostatistics, Univ of North Texas Health Science Center, Fort Worth, TX, ksingh@hsc.unt.edu
‡School Electrical Engineering and Computer Science, Washington State University, Pullman, WA, cook@eecs.unt.edu

***Abstract*—Analysis of Google Influenza-like-illness (ILI) search queries has shown a strongly correlated pattern with Center for Disease Control and Prevention seasonal ILI reporting data. Web and social media (WSM) provide another resource to detect increases in ILI. This paper evaluates trends in blog posts that discuss Influenza. The results of the analysis show Influenza-related blogging trends have a significant correlation with the beginning of US Fall 2008 flu season. We also identify WSM Influenza-related communities that share flu-postings which could broker or disseminate information in the case of a severe outbreak or Influenza epidemic.**

I. Introduction

Influenza diagnosis based solely on the presentation of symptoms is limited as these symptoms may be associated with many other diseases. Serologic and antigen tests require that a patient with Influenza-like-illness (ILI) be examined by a physician who can either conduct a rapid diagnosis test or take blood samples laboratory testing. This suggests that many cases of Influenza remain undiagnosed. While the presence of Influenza in an individual can be confirmed through specific diagnostic tests, the Influenza prevalence in the population at any given time is unknown and can only be estimated. In the past, such estimates have relied solely on the extrapolation of diagnosed cases, making it difficult to identify the various phases of seasonal Influenza, or the identification of a more serious manifestation of a flu epidemic.

Web and social media (WSM) provide a resource to detect increases in ILI. This paper evaluates blog posts that discuss Influenza, analysis show a significant correlation with the beginning of US Fall 2008 flu season. A well defined response strategy to an outbreak may make use of WSM to reduce population and human impact of the disease. We suggest a possible response that identifies WSM Influenza-related communities that share flu-postings. These community or crowd sources could broker and disseminate important intervention information in the case of a infectious disease outbreak. The flowchart in Fig. 1 visually describes this approach to detecting and responding to Influenza trends.

We briefly discuss a history of infectious disease outbreaks and recent approaches in online Public Health surveillance of Influenza and the value of social community is discussed with regards to outbreak responses. Next, the data set used in our analysis is presented and the methodology for information extraction and trend analysis is outlined posting trends. Through discovery and verification of trends in Influenza related blogs, we verify a correlation to Center for Disease Control and Prevention (CDC) Influenza-like-illness patient reporting at sentinel healthcare providers. Additionally, categories, frequency and Influenza-post persistence qualitatively assist ILI trend identification in blogs. Strongly connected communities are evaluated and influential bloggers identified that should be part of an WSM outbreak response. This paper concludes with approaches to expand ILI-trend identification and an approach to create an integrated Public Health and WSM community intervention campaign.

A. Background

Epidemics of infectious diseases have plagued humankind since historical times. There are accounts of epidemics dating back to the times of Hippocrates (459 377 B.C.) and the ancient Greeks [1]. Fourteenth century Europe lost a quarter of its 100 million people to Black Death. The fall of the Aztecs empire in 1521 was due to smallpox that eradicated half of its 3 $\frac{1}{2}$ million population. The pandemic Influenza of 1918 caused over 20 million excess deaths in 12 months. More recently, the severe acute respiratory syndrome (SARS) outbreak of 2003 highlighted the rapid spread of an epidemic at the global level. The outbreak, emanating from a small Guangzhou province in China, spread around the world requiring a concerted response from public health administrations around the world and the World Health Organization (WHO) to curtail the epidemic [10]. The WHO and (CDC) [4] actively engage in worldwide surveillance of infectious diseases, and prioritize prevention and control measures at the root cause of epidemics.

The pervasiveness and ubiquity of the Internet and World Wide Web resources provide individuals with access to many information sources that facilitate self-diagnosis; one can combine specific disease symptoms to form search queries. The result of such queries often lead to sites that may help diagnose the illness and offer medical advice. Recently, Google has addressed this issue by capturing the keywords of queries and identifying specific searches that involve search terms that indicate Influenza-like-illness [8]. Other published research on Influenza Internet surveillance include search "advertisement click-through" [6], using a set of Yahoo search queries containing the words flu or Influenza [15], and health website access logs [12].

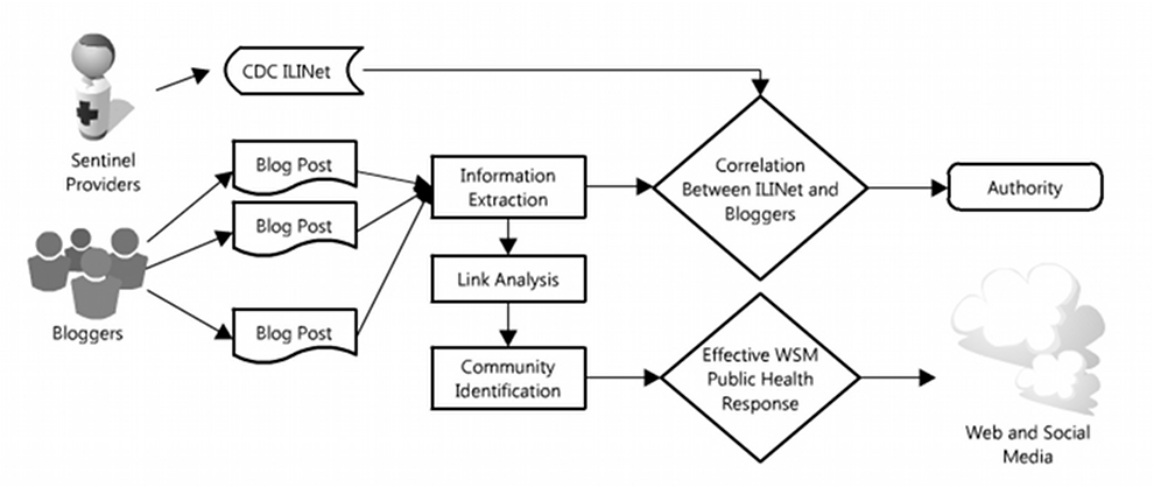
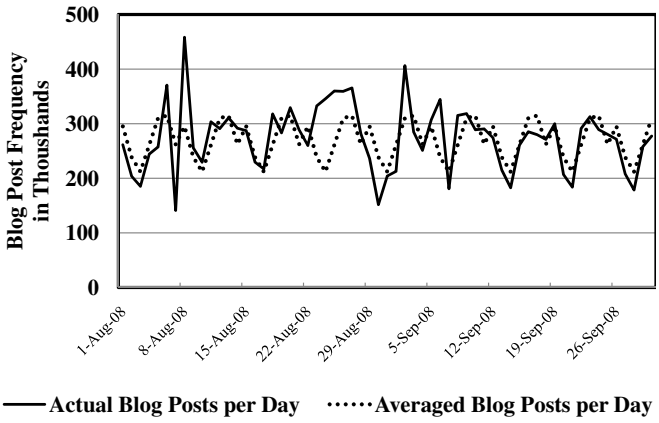


Fig. 1. Methodology to monitor Influenza-like-illness in social media and to identify possible web and social media communities to participate in a Public Health response

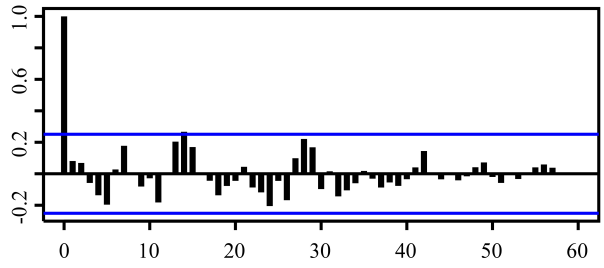
II. Data and Methodology

Blog data used for the Influenza trend identification is provided by *Spinn3r* [11]. The data contains 44 million UTF-8 encoded posts(items) collected from 1-August to 1-October 2008, its physical size is 196 GB. Blogs are classified into a ranked tier structure with the tiers defined by how successful it is at creating and participating in memes on the Internet (`http://tailrank.com`). There are 13 rank tiers and a 14^{th} tier with no rank value; due to processing constraints this paper uses blog items between 1-August and 30-September 2008 and are ranked in the first to twelfth tier; 8,654,874 blog items in August and 8,027,370 blog items in September 2008. We assume the tier range for this task is sufficient by the Tailrank classification definition. From these, English language blog items are extracted by matching the `<dc:lang>` tag to en or by U tagged items whose `<description>` (html content) is encoded in ascii. This paper defines the *blog-world* to be English language tier 1 to 12 blog posts. We also consider the following terms to be equivalent: blog post & blog item and blogger & blog site. Indexing, parsing and link extraction code was written in Python, parallelized using pyMPI and executed on an eight node cluster, each node has 32 GB of memory and two 2.66GHz Quad Core Xeon processors [16], [14].

To facilitate anomaly detection, one can compare baseline blog-world trends to specific topic (e.g. Influenza) trends. First, day of week posting trends are identified, the per day frequency of blog posts is calculated and plotted in Figure 2a. The blog post date is taken from the publication date tag found in the item metadata. During August and September 2008, a clear seven day *period* in posting frequency is observed. Posting is more frequent during the middle of the week and less prevalent on the weekends, defining a seven day cycle that begins Sunday and ends on Saturday. The frequency for each of the seven days in a week, is averaged over the two month period and superimposed in Figure 2a. A seven day posting period is present in these averages as well; most bloggers author content weekdays and less on the weekends. Time spent with friends and family, relaxing and not working may explain



(a) Actual vs Average Blog-World Posts Per Day of Week



(b) Blog-World Post Frequency Autocorrelation. The X-axis is labeled by the number of days since 1-August 2008 and the autocorrelation value is plotted on the Y-axis.

Fig. 2. August and September 2008 Blog-World Posts. The *blog-world* is defined to be English language blog posts having a Tailrank value between one and twelve.

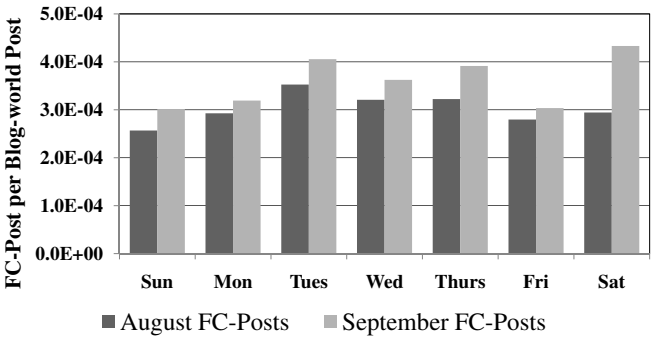
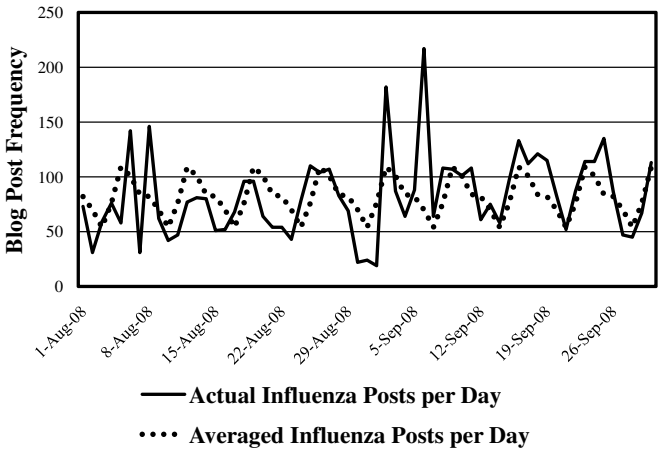
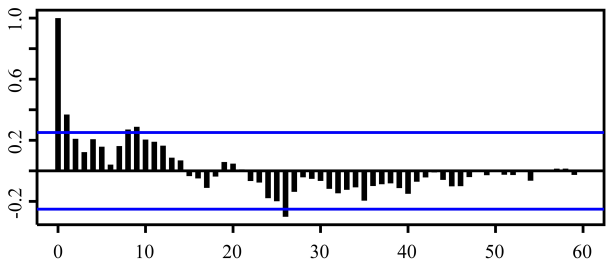


Fig. 3. Average FC-posts Per Day of Week; normalized by corresponding average day of week blog-world post count



(a) Actual vs Average FC-Posts Per Day of Week



(b) FC-post Frequency Autocorrelation. The X-axis is labeled by the number of days since 1-August 2008 and the autocorrelation value is plotted on the Y-axis.

Fig. 4. August and September 2008 FC-post. A *FC-post* is an English language, tier one to twelve blog post containing the term *Influenza* or *flu*.

the low number of weekend posts. In addition to detecting anomalies by posting variance, changes in the period length, beyond what may be expected by randomness, flag anomalies worthy of further investigation.

To verify the seven day posting cycle is not a random artifact, the autocorrelation function (ACF) is relevant for this task. The ACF facilitates identification of period length in a time series and determines if the cycle is non-random; the ACF is most commonly used in signal processing [2]. Autocorrelation measures the correlation between values of the same variable at times X_i and X_{i+k}. Given measurements (posts per day), $Y_1, Y_2, \ldots Y_N$ at time $X_1, X_2, \ldots, X_N$, Equation 1 defines the lag k autocorrelation function (X measurements are assumed to be taken at the same interval). Figure 2b displays the blog-world data ACF[1]. Since the first period of the ACF remains inside the confidence interval, the blog-world post frequency *period* is not random. In addition, all ACF time-lag steps lie in the 95% confidence intervals, proving the seven day period in posting frequency is not random. This technique is valuable when monitoring shifts in posting frequency that occur outside of what is possible by randomness.

$$r_k = \sum_{i=1}^{N-k} \frac{(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \quad (1)$$

III. Analysis of Trends in Influenza Blogs

In our analysis, we extract English language items from the blog-world index when a lexical match exists to the terms *Influenza* and *flu* anywhere in its content (misspellings and synonyms are not included). The blog items are grouped by month, week (Sunday to Saturday) and by day of week. The extracted blog items containing Influenza keywords are termed flu-content posts or *FC-posts*.

A. Monitoring Influenza Trends

Monitoring FC-posts for sudden shifts in posting frequency, beyond what is randomly possible, could identify increased ILI. A seven day posting period is also present in Influenza related blog items, with posting frequency higher weekdays and lower on weekends. The day of week FC-post frequency is averaged for August and September and is normalized by the corresponding day of week average for all blog-world posts, this data is plotted in Fig. 3. A slight posting frequency period can be observed for the month of August in Fig. 3; however, the posting variation is not significant. The first 30 lags (days) in the ACF verifies (Fig. 4b) August's slight seven day FC-posting period is random and any variation in the period cannot be differentiated from random noise. Given this information, we can assume the August level of ILI is stable. However, starting in September (historically, the month preceding flu season) the daily FC-post average increases significantly. Referring to the Fig 4b, a non-random seven day period can be observed beginning 28-August. Throughout September 2008, a clear weekly FC-post pattern emerges and is is not a random pattern, within 95% confidence.

The nearly three-fold Saturday FC-post increase in September compared to August is a significant anomaly, the ACF verifies that the shift is not a random occurrence. On September 6, the number of ILI-posts are nearly triple the average ILI-posts for a Saturday and double the total average post per day (see Fig. 4a). One might speculate the return to school

[1]The ACF at time 0 is unity because there is no previous data to correlate and thus cannot be differentiated from random.

of children, adolescents and college students contributes to the extraordinary increase of FC-posts. Perhaps, individuals contract an ILI in the confines of school or college and bring it home to their families; alternatively, students and other bloggers could be conversing about the flu and elevate the flu items present in the blog-world. September 1st is Labor day, an American holiday signaling the end of Summer break in many regions of the US and the beginning of Fall school. We see a dramatic decrease in posts September 1, both in the blog-world and FC-posts. We posit that most bloggers are spending time away from their computers, engaging in face-to-face time with family and friends, barbecuing and enjoying the calm before a grueling trudge to the end of the year. Also, no one likes to stay home on a holiday, even if they are sick.

B. FC-post Trends

FC-post trends can be monitored using the social media mining methodology presented in this paper. This methodology facilitates identification of outbreaks or increases of Influenza infection in the population. This paper's most significant finding is a strong correlation between the frequency of FC-posts per week and Center for Disease Control and Prevention Influenza-like-illness surveillance data. Qualitative assessment of category tags, prevalence of FC-posts on a blog site, and persistent posting of flu related posts also suggest ILI trends.

1) Correlation to CDC ILINet: We posit that the increase in September 2008 blog-world flu-posts correlate with an increase of ILI and start of flu-season, to verify this statement we compare our data to Center for Disease Control and Prevention surveillance reports from sentinel healthcare providers. The CDC website states the Outpatient Influenza-like-illness Surveillance Network (ILINet) consists of about 2,400 healthcare providers in 50 states reporting approximately 16 million patient visits each year. Each report data to CDC on the total number of patients seen and the number of those patients with Influenza-like-illness (ILI) by age group. For this system, ILI is defined as fever (temperature of 100F [37.8C] or greater) and a cough and/or a sore throat in the absence of a known cause other than Influenza [4].

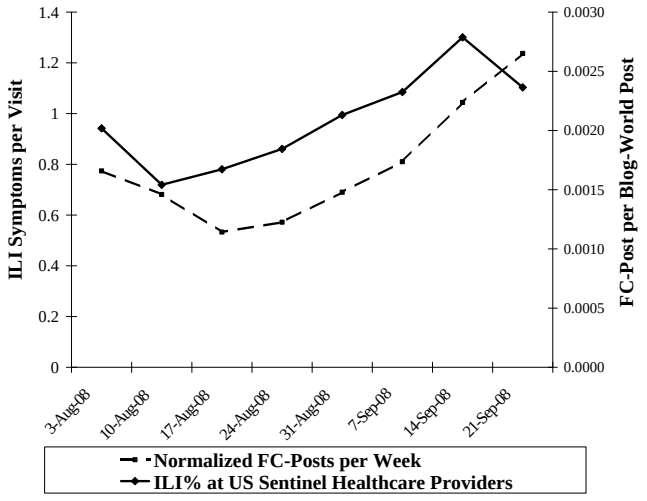


Fig. 5. CDC ILINet vs Normalized FC-post Frequency per Week. Each FC-posts per week data point is normalized by the corresponding blog-world posts per week count

The CDC ILINet surveillance and FC-post per week data are plotted in Fig. 5. CDC Influenza-like-illness symptoms per visit at sentinel US healthcare providers labels the primary Y-axis. The secondary Y-axis marks the FC-post per week frequency normalized by the corresponding blog-world week post count. Correlation between the two data series measured with a Pearson correlation coefficient, r. To prove our hypothesis that a correlation between CDC ILINet reports and social media mined FC-post frequency, Pearson's correlation coefficient is evaluated between the two data series. The Pearson correlation evaluates to unity if the two data series are exactly matching, $r = 1$. If no correlation exists between the data series, the Pearson correlation evaluates to zero, $r = 0$. In our analysis, the eight ILI and FC-post data points correlate strongly with a high Pearson correlation, $r = 0.767$, and the correlation is significant with 95% confidence.

C. FC-post Types

Once FC-posts have been extracted, one can further monitor Influenza outbreaks by evaluating the perspective of blog authors. Bloggers having a direct knowledge of Influenza infection are more valuable to disease surveillance than those who author objective or opinion items. Bloggers who persistently author FC-posts are less likely to be infected with Influenza and more likely to be writing about Avian Influenza (Bird Flu).

Identifying the perspective of Influenza keyword posts facilitates determining its contribution to disease surveillance, three author perspectives are identified. A FC-post is either a self-identification of having ILI symptoms, secondhand (or by proxy) of another individual having ILI or the post is an opinion or objective article containing ILI keywords. Secondhand knowledge can be writing about a friend, schoolmate, family-member or co-worker but a blogger could also post details on famous individual such as a sports player. The season opening of American football coincides with the data and many FC-posts identify athletes who are unable to play because of an ILI, see the excerpt below. The following post excerpts demonstrate each FC-post type

· **Self Identified** : "What began as an irritating cold became what I think might be the flu last night. I woke up in bed around three this morning with sore muscles, congested lungs/nose and chills running throughout my body."

· **Secondhand** : "According to ESPN.com, Ravens quarterback Troy Smith has lost 'a considerable amount of weight' while being hospitalized with tonsillitis and flu-like symptoms. Smith and veteran Kyle Boller likely won't play in Sunday's season opener, leaving the workload to rookie Joe Flacco and Joey Harrington, who was signed Monday."

· **Objective (or Opinion) Article** : "Domesticated birds may become infected with avian Influenza virus through

TABLE I
TOP FIVE FLU CONTENT BLOGGERS

Aug Blogs	
124	http://birdcauseflu.com
96	http://crofsblogs.typepad.com/h5n1
59	http://birdflubreakingnews.com
49	http://medblogs.org
34	http://afludiary.blogspot.com
September Blogs	
82	http://birdcauseflu.com
50	http://crofsblogs.typepad.com/h5n1
34	http://afludiary.blogspot.com
33	http://medblogs.org
26	http://bird-site.com

direct contact with infected waterfowl or other infected poultry, or through contact with surfaces or materials like that of water or feed that have been contaminated with the virus."

How often or how persistently bloggers author FC-posts indicate trends as well, a blog-site that has FC-posts for a limited time is more likely to be a first or secondhand experience.

Assigning each FC-post to a set, constrained by month, is another approach to identifying anomalies. Figure 6 demonstrates only a small portion of bloggers author "Influenza" posts each month, whereas the majority of flu bloggers author content only once during the two months. The 1,167 and 1,927 blog posts in August and September, respectively, mark a nearly two fold increase in the number of singleton bloggers writing about Influenza. We do not account for the number of bloggers who have an ILI and author content during both months, or if they get sick one month and write secondhand knowledge the other month. Accounting for these posters would only increase the number of singleton Influenza posters, not negatively effect the trend surveillance.

Monitoring self-identification and secondhand FC-post trends can mark increases in ILI. It can be said that bloggers that post often about Influenza are more likely to a) be an authority on Influenza (perhaps not an expert though) where its readers find information on Influenza or b) the blogger is frequently sick with Influenza. The five most frequent FC-post bloggers in August and September (Table I) demonstrate the hypothesis frequent flu bloggers are news / opinion oriented.

1) Categories: Categories for each blog site are extracted from category metadata associated with each blog item. If a blogger uses a category in more than one post, the category is only counted once. Table II lists the top 35 categories and how often they appear. Avian flu and the H5N1 virus postings

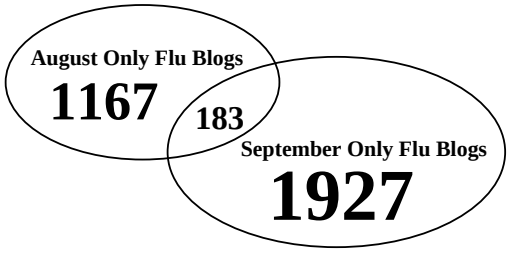


Fig. 6. The number of bloggers who author FC-posts in August only, September only, or in both August and September

TABLE II
BLOG CATEGORY OCCURRENCE PER MONTH

August 2008		September 2008	
Blogs	Category	Blogs	Category
120	AVIAN FLU	103	AVIAN FLU
58	HEALTH	67	HEALTH
55	NEWS	52	BIRD FLU SYMPTOMS
54	BIRD FLU SYMPTOMS	47	NEWS
45	GENERAL	46	LIFE
39	LIFE	42	POLITICS
36	MUSIC	40	GENERAL
25	BLOG	33	BUSINESS
21	FAMILY	30	MUSIC
21	MOVIES	26	BOOKS
19	POLITICS	22	SICK
18	FLU	22	INFLUENZA
18	TRAVEL BLOGS	21	MEME
17	BUSINESS	19	NO-TAG
16	ENTERTAINMENT	17	ENTERTAINMENT
16	NEWS AND POLITICS	16	SPORTS
16	PODCASTS	16	FAMILY
16	FLU / COLD / SARS	15	BLOG
15	BLOGGING	15	FLU
15	SPORTS	15	TRAVEL BLOGS
14	MEME	14	LATEST NFL HEADLINES
14	PERSONAL	14	FLU / COLD / SARS
14	FOOD	13	HUMOR
14	PROPHECY	13	CULTURE
13	HEALTH AND WELLNESS	13	NEWS AND POLITICS
13	PODCAST	13	MOVIES
13	AVIAN	13	UK
12	SICK	12	TECHNOLOGY
12	UK	12	C. ELEGANS
11	TECHNOLOGY	12	FOOD

remain nearly constant in August and September. However, bloggers tagged with a "SICK" category double between the months, further verifying our ILI trend increase in September. Back to our reference to American football athletes contracting Influenza, we see a greater number of "SPORTS" categories and the "LATEST NFL HEADLINES" category has appeared on 14 FC-post blogs.

IV. RESPONSE STRATEGY IN "FLU" BLOG COMMUNITIES

Development of accurate predictive health related outcome tools is imperative as the United States and other nations strive to increase individual and community health while reducing economic burden. Effective intervention strategies can be delivered to targeted communities through large-scale social media data-mining, demographic population analysis and predictive modeling of risk behavior. WSM communities will play a vital role in any Public Health response to an outbreak. Influential bloggers can disseminate and broker response strategies/interventions in their blog communities, the bloggers could be first responders to a disease outbreak, in an information sense. Their readers will hopefully trigger an information cascade, spreading the response, to vaccinate, quarantine, school closings etc. Remiss in our data is more recent evolutions of WSM community such as micro-blogging (the number of Twitter accounts grew by 900% in 2008), and wiki-style communities which many times are "gated" and not indexed in shallow web-crawls.

We take an intuitive and simple definition of WSM community and identify possible first responder bloggers by link analysis. Blog ranking enhances the idea these communities can disseminate information as part of a broader Public Health response triggered by anomalies in ILINet and WSM surveillance. Community is defined similarly to Flake, Lawrence

and Giles, we consider strong flu post ties, links only occur between FC-post bloggers [7]. Links from a non FC-post to an FC-post and vice versa are not defined in this community definition. Although considerably less costly than a main stream media campaign, a WSM targeted response must be cost-effective and optimized to achieve maximum strategy penetration. Any blogger participating in Public Health campaign needs to have influence in its community, and the ability to disseminate information to other WSM. Closeness and betweenness centrality measures, and Google's PageRank (eigenvector centrality) will rank Influenza community blog sites in order to target key actors.

Wasserman and Faust state closeness can be productive in communicating information to other actors. It is defined in Eqn. 2 as the average shortest paths or geodesics distance from actor v and all reachable actors in [17] :

$$C_{C_v} = \frac{\sum_{t \in V \setminus v} d_G(v,t)}{n-1} \tag{2}$$

Betweenness centrality (Eqn. 3) measures interpersonal influence, specifically an blog is central if it lies between other blogs on their geodesics - the blog is "between" many others, where g_{jk} is the number of geodesics linking blog j and blog k [17] :

$$C_{B_v} = \sum_{j<k} \frac{g_{jk}(n_i v,t)}{g_{jk}} \tag{3}$$

Page Rank is an example of eigenvector centrality and measures the importance of a node by assuming links from more central nodes contribute more to its ranking than less central nodes [3]. Let d be a damping factor (usually 0.85), p_n are the pages, $M(p_i)$ is the set of pages linking to p_n and $L(p_j)$ is the outlink counts on page p_j :

$$R_{p_n} = \frac{1-d}{N} + d \sum_{p_j \in M(p_n)} \frac{PR(p_j)}{L(p_j)} \tag{4}$$

Three centrality measures are calculated for the largest FC-post community member blog sites. August 2008's largest community (Table III) has 37 member blogs, and September 2008's largest community (Table III) has 14 member blogs. In developing an Public Health WSM response plan, each type of centrality better classifies how a blog site will influence and disseminate pertinent information. Blogs with high *betweenness* could broker information between bridged communities, synchronizing knowledge. Blogs with larger *closeness* and *PageRank* values can quickly disseminate outbreak response strategies.

V. Discussion

The fusion of Complex Network Analysis, Health Informatics and Computational Linguistics to unrelated disciplines such as Sociology, Economics and Public Health is broadening interdisciplinary participation of scientific collaborative research. New hybrid approaches composed of integral theories and practices from these areas creates a synergy enabling discovery and understanding on how communications, information, networks and community structure effect population health.

Public health professionals often have limited budgets and resources must be specifically tailored to achieve maximum results. The utilization of computational social networking tools would allow for those within the public health industry to anticipate the impact of community specific predictions, and tailor awareness, educational, intervention, and prophylactic programs for the greatest impact within their population. Consider a network of social interactions and communication links whose nodes are the family unit, local health departments, clinics, Web and Social Media (WSM) and main stream media (MSM); modeling what-if scenarios using simulation frameworks, WSM and Public Health data analysis will increase Public Health response times and decrease negative impacts of a serious Influenza outbreak.

The trend spotting and response framework presented could be improved by extending our identification and classification of WSM for Influenza outbreaks. Ginsberg et al in the Feb 2009 Nature article automatically determine search queries effective in modeling seasonal Influenza [8]. Expanding blog

TABLE III
August and September 2008 EN-language *Flu* Community

PR	URL	CLOSE	BTWN
Largest Component of Flu-Linked Blogs : August 2008			
0.154	http://crofsblogs.typepad.com/h5n1/	0.514	0.715
0.099	http://birdflumonitor.com	0.33	0.263
0.083	http://afludiary.blogspot.com	0.424	0.3
0.069	http://www.eurekalert.org/	0.327	0.262
0.065	http://newfluwiki2.com/frontPage.do	0.424	0.374
0.049	http://fluwikie2.com	0.391	0.056
0.04	http://birdflujourney.typepad.com	0.419	0.256
0.032	http://www.eurekalert.org/pubnews.php	0.31	0.11
0.029	http://medicalnewstoday.com	0.391	0.048
0.027	http://fromthestyx.wordpress.com	0.25	0
0.018	http://Influenzapandemic.blogspot.com	0.288	0.001
0.018	http://allafrica.com/eastafrica	0.305	0.006
0.017	http://www.unicef.org/	0.333	0.021
0.016	http://sciencewriter.spaces.live.com	0.25	0
0.016	http://hatingautism.blogspot.com	0.25	0
0.016	http://advancednano.blogspot.com	0.25	0
0.016	http://infowars.org	0.25	0
0.015	http://channelnewsasia.com	0.3	0
0.015	http://www.japantoday.com/	0.367	0
0.015	http://allafrica.com/westafrica	0.343	0
0.014	http://shakesville.com	0.248	0
0.014	http://iith096.blogspot.com	0.248	0
0.014	http://ykalaska.wordpress.com	0.248	0
0.014	http://www.echojournal.org/	0.248	0
0.014	http://www.geeknewscentral.com/podcasts/	0.248	0
0.013	http://smarteconomy.typepad.com/smart_economy/	0.238	0
0.013	http://birdsearcher.com	0.238	0
0.01	http://crofsblogs.typepad.com	0.343	0
0.01	http://flupatrol.com	0.343	0
0.01	http://www.canada.com/index.html	0.343	0
0.01	http://dailykos.com/section/diary	0.343	0
0.01	http://allafrica.com/nigeria	0.343	0
0.01	http://californiaprogressreport.com	0.3	0
0.01	http://ww.tpmcafe.com	0.343	0
0.01	http://allafrica.com/afdb/blogs/	0.343	0
0.01	http://www.msnbc.msn.com/id/3032076/	0.283	0
0.009	http://plosone.org/home.action	0.3	0
Largest Component of Flu-Linked Blogs : September 2008			
0.159	http://newfluwiki2.com/frontPage.do	0.542	0.359
0.149	http://afludiary.blogspot.com	0.481	0.308
0.132	http://birdflumonitor.com	0.351	0.154
0.131	http://crofsblogs.typepad.com/h5n1/	0.619	0.562
0.129	http://fluwikie2.com	0.52	0.295
0.065	http://medicalnewstoday.com	0.433	0.092
0.031	http://fiercebiotech.com	0.265	0
0.029	http://flupatrol.com	0.394	0
0.029	http://Influenzapandemic.blogspot.com	0.394	0
0.029	http://www.pressgazette.co.uk/	0.394	0
0.029	http://fijitimes.com	0.351	0
0.029	http://www.earthtimes.org/	0.351	0
0.028	http://usatoday.com/news/health/front.htm	0.361	0
0.028	http://southernstudies.org/facingsouth/	0.361	0

post extraction to include keywords from Table IV may enable greater accuracy in detecting FC-post trends. The methodology presented can be extended to other diseases. The CDC produces an annual list of notifiable diseases, given data in a similar format to ILINet surveillance reports, this framework can be shifted to monitor their trends. Future work will take advantage of WSM location tagging, trends will be collocated with geographic regions creating opportunities for decisive disease intervention.

TABLE IV
INFLUENZA KEYWORD EXPANSION

Google Search Queries Reported in [8]
Influenza complication
Cold/flu remedy
General Influenza symptoms
Term for Influenza
Specific Influenza symptom
Symptoms of an Influenza complication
Antibiotic medication
General Influenza remedies
Symptoms of a related disease
Antiviral medication
Related disease
Unrelated to Influenza

Information in social media exists, flows and interacts between agents with those agents possibly interacting or communicating in multiple modes. General models exist simulating innovation diffusion and spread of contagion in social networks and also information transmission and cascades in communication networks. Empirical analysis of multiple WSM data (e.g. micro-blogs, blog comments) integrated with Federal, State and Local Public Health communication resources enable effective responses [5], [9], [13].

VI. CONCLUSION

Web and social media (WSM) provide a resource to detect increases in ILI. We presented a method which evaluates blog posts containing keywords Influenza or flu and the results from analysis show a significant correlation with the beginning of US Fall 2008. This paper's most significant finding is a significant and strong correlation between the frequency of FC-posts per week and Center for Disease Control and Prevention Influenza-like-illness surveillance data. Additionally, categories, frequency and constancy qualitatively assist ILI trend identification in blogs. Strongly connected communities are evaluated and influential bloggers identified that should be part of an WSM outbreak response. These community or crowd sources could broker and disseminate important intervention information in the case of a infectious disease outbreak.

ACKNOWLEDGEMENT

We would like to thank the National Science Foundation (NSF) for support under grant NSF IIS-0505819. The contents of this publication are the responsibility of the authors and do not necessarily represent the official views of the NSF.

REFERENCES

[1] N. Bailey, *The Mathematical Theory of Epidemics*, 1957. [Online]. Available: http://www.google.com/search?client=safari&rls=en-us&q=The+Mathematical+Theory+of+Epidemics&ie=UTF-8&oe=UTF-8

[2] R. N. Bracewell, "The autocorrelation function," *The Fourier transform and its applications*, pp. 40–45, Jan 1965. [Online]. Available: http://books.google.com/books?id=undFOgAACAAJ&printsec=frontcover

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, Jan 1998. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S016975529800110X

[4] CDC-Website, "Influenza surveillance reports," Website, 2009, http://www.cdc.gov/flu/weekly/fluactivity.htm. [Online]. Available: http://www.cdc.gov/flu/weekly/fluactivity.htm

[5] M. Chwe, "Structure and strategy in collective action," *American Journal of Sociology*, Jan 1999.

[6] G. Eysenbach, "Infodemiology: tracking flu-related searches on the web for syndromic surveillance," *AMIA Annual Symposium proceedings*, pp. 244–8, Jan 2006.

[7] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of web communities," in *In Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 2000, pp. 150–160.

[8] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–4, Feb 2009.

[9] M. Granovetter, "The impact of social structure on economic outcomes," *Journal of Economic Perspectives*, Jan 2005. [Online]. Available: http://www.jstor.org/stable/4134991

[10] D. Heymann and G. Rodier, "Global surveillance, national surveillance, and sars," *Emerging Infectious Diseases*, vol. 10, no. 2, Feb 2004. [Online]. Available: http://www.google.com/search?client=safari&rls=en-us&q=Global+Surveillance,+National+Surveillance,+and+SARS&ie=UTF-8&oe=UTF-8

[11] ICWSM, "Icwsm 2009 spinn3r dataset," in *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, May 2009.

[12] H. A. Johnson, M. M. Wagner, W. R. Hogan, W. Chapman, R. T. Olszewski, J. Dowling, and G. Barnas, "Analysis of web access logs for surveillance of influenza," *Studies in health technology and informatics*, vol. 107, no. Pt 2, pp. 1202–6, Jan 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15361003?dopt=abstract

[13] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the ninth ACM SIGKDD*, Jan 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=956750.956769&type=series

[14] P. Miller, "pympi–an introduction to parallel python using mpi," *Livermore National Laboratories*, Jan 2002. [Online]. Available: https://computing.llnl.gov/code/pdf/pyMPI.pdf

[15] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson, "Using internet searches for influenza surveillance," *Clin Infect Dis*, vol. 47, no. 11, pp. 1443–8, Dec 2008.

[16] G. V. Rossum and F. Drake, "Python language reference," *Network Theory Ltd*, Jan 2003. [Online]. Available: http://www.altaway.com/resources/python/reference.pdf

[17] S. Wasserman and K. Faust, "Social network analysis: methods and applications," p. 825, Jan 1994. [Online]. Available: http://books.google.com/books?id=CAm2DpIqRUIC&printsec=frontcover