Integration and Visualization Public Health Dashboard: The *medi+board* Pilot Project

Patty Kostkova, Stephan Garbin, Justin Moser, Wendy Pan
Department of Computer Science
University College London
Gower Street
London, UK
{P.Kostkova, Justin.Moser.13, Stephan.Garbin.13, Wendy.Pan.13}@ucl.ac.uk

ABSTRACT

Traditional public health surveillance systems would benefit from integration with knowledge created by new situation-aware realtime signals from social media, online searches, mobile/sensor networks and citizens' participatory surveillance systems. However, the challenge of threat validation, cross-verification and information integration for risk assessment has so far been largely untackled.

In this paper, we propose a new system, *medi+board*, monitoring epidemic intelligence sources *and* traditional case-based surveillance to better automate early warning, cross-validation of signals for outbreak detection and visualization of results on an interactive dashboard. This enables public health professionals to see all essential information at a glance. Modular and configurable to any 'event' defined by public health experts, *medi+board* scans multiple data sources, detects changing patterns and uses a configurable analysis module for signal detection to identify a threat. These can be validated by an analysis module and correlated with other sources to assess the reliability of the event classified as the *reliability coefficient* which is a real number between zero and one. Events are reported and visualized on the *medi+board* dashboard which integrates all information sources and can be navigated by a timescale widget.

Simulation with three datasets from the swine flu 2009 pandemic (HPA surveillance, Google news, Twitter) demonstrates the potential of *medi+board* to automate data processing and visualization to assist public health experts in decision making on control and response measures.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences - *Medical information systems.*

General Terms

Experimentation, Algorithms.

Keywords

Epidemic Intelligence, Outbreak Detection, Cross-validation, real-time data scanning, Dashboard.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579276

1. INTRODUCTION

Prevention, management and control of infectious diseases remain on the forefront of public health activities. The importance of this role for citizens worldwide has been recently highlighted by outbreaks such as SARS in 2003 and swine flu in 2009. However, the potential of an increasingly growing amount of information on the Internet for digital epidemiology has been substantially increased with the arrival of social media and Web 2.0 platforms enabling near real-time event tracking. For example, this applies to large population movements which can be monitored by exploiting geographic and spatiotemporal tags.

Digital epidemiology harvesting digital data sources for public health purposes brings great potentials and new challenges [1], and creates new possibilities for the use of Big Data [2]. Complementing traditional case-based microbiological laboratory reports and syndromic surveillance, *event-based* surveillance is monitoring unstructured events, such as news, and has been a significant component of public health early warning and response over the last decade (GPHIN, MedISys).

Further, the roadmap for digital epidemiology incorporating new data sources was recently outlined [3] discussing a set of data streams ranging from traditional surveillance datasets, lab-confirmation reports and citizens participatory surveillance systems (such as Influenzanet) to real-time situation-aware geo-located information.

In this paper, we first turn to an examination of the background in section 2. In section three, we present the *medi+board* public health dashboard vision and elaborate on the design of its infrastructure in section 4. Section 5 brings implementation details, followed by a demo conducted with data from the swine flu outbreak of 2009 in section 6. Future work is presented in section 7 while section 8 concludes.

2. BACKGROUND

Recently, the role of public health and the importance of utilizing digital information significantly increased due to global travel and the emergence of new diseases (such as SARS). Traditionally, case-based and syndromic surveillance relies on national surveillance systems with established bottom-up reporting processes from local through regional to national and international levels. However, the process of reporting, collating and analyzing data normally takes several weeks which hinders a targeted response in the early stages of a pandemic. The emerging discipline of *Epidemic Intelligence* (EI) is made possible by geographic and spatiotemporal tags found in digital communication and could help to overcome this limitation. However, current EI systems typically focus on one particular source which makes their application somewhat limited (BioCaster, Argus, GPIHN, HealthMap, MedISys, ProMED-mail, Puls). This was identified by a comparative study on the detection of A/H5N1 Influenza Events [4] which also highlighted the need for *"more efficient synergies and cross-fertilization of knowledge and information"*, which is what we are concerned with.

The roadmap for a digital disease surveillance dashboard incorporating new data sources was recently outlined [3] highlighting six types of data sources for EI: *news/online media*, *digital traces*, *Pro-Med*, *labs/clinical reports*, *participatory systems* and *social media.*

For over a decade, *online media* has been the prime source for epidemic intelligence. Tools such as Global Public Health Intelligence Network (GPHIN) [5], developed by Health Canada and in use by the WHO, and Medisys[1] gather news from global media to identify disease outbreaks threats using multi-lingual natural language processing and appropriately weighted set of keywords, categories and taxonomies [6,7]. An unstructured event-based reports from GPHIN [5], HealthMap [8] and EpiSPIDER[2], were analyzed for global infection disease surveillance and future development outlined in Keller et al [9]. News are however not suitable for early warning systems as it usually takes several days for an event to be reported. Secondly, not all countries exhibit free-press coverage, making official news unreliable and causing significant delays. This increases the importance to rely on other sources such as social media in these cases.

Recently, *digital traces* have become essential signal sources including search keywords, loyalty cards, sensor networks, drugs purchases and mobile phone data. Regrettably, these systems typically rely on non-publicly available, company internal datasets and are thus are not easily available for research. Google's Flu Trends [10] is an example of this kind of proprietary work which provides no means for verification or direct comparison. Ginsberg et al [11] illustrated an automated method for defining ILI-related keywords without prior knowledge of influenza. A similar study, investigating search keywords and online behaviour by infection experts was conducted by Wiseman et al identifying information needs during major outbreaks from weblog searches [12, 13].

Thirdly, the email-based system *ProMED-mail*[3] has been a long-established informal source of emergencies discussed by infectious disease professionals. It's 'informality' stems from the fact that as human moderated data source it is subject to bias and has a comparatively low coverage.

Labs and clinical reports are traditionally regarded as the backbone of surveillance systems. Microbiological laboratories contribute to surveillance by confirmation of unusual disease patterns and specimen (albeit at the expense of timeliness).

Participatory systems (web-based or mobile) require pro-active participation in terms of regularly sharing disease symptoms collected in a structured format (examples are the above mentioned multilingual EpiWorks project Influenzanet[4]). Unlike the popular social media platforms, participatory systems typically limit submissions to a set of symptoms, thus sacrificing coverage at the expense of making user contributions easier.

[1] http://medusa.jrc.it/medisys/homeedition/en/home.html

[2] http://www.epispider.org/

[3] http://www.promedmail.org/

[4] http://www.influenzanet.eu/

Social media sources have revolutionized the speed and timeliness of EI. Information posted on twitter describes real time activity unlike queries collected by search engines. Twitter can be used to both track [14, 15] and even predict [16] the spread of infectious diseases as we demonstrated in our previous study. Lampos [17] used their technique of supervised learning for 'nowcasting' events by exploring geo-located Twitter signal. ILI were tracked and correlated with CDC surveillance data also by Culotta [18] and a dengue fever was tracked using Twitter in Brazil by Gomide et al [19]. Recently, Salathe et al illustrated the role of digital epidemiology and Twitter for understanding the new strain of Influenza A (H7N9) and the coronavirus (MERS-CoV) [20]. Signorini et al evaluated user sentiment during the swine flu outbreak in the US and influenza like illness (ILI) reported disease levels [21].

3. *medi+board* PUBLIC HEALTH DASHBOARD

In this section, we present the overview of the integrated digital public health dashboard and the infrastructure required for data-mining, threat detection, verification, correlation of threats to create 'events' and subsequent reporting.

The framework illustrated in Figure 1 which was defined in [3] depicts processes and components required for automated monitoring across multiple realtime data channels.

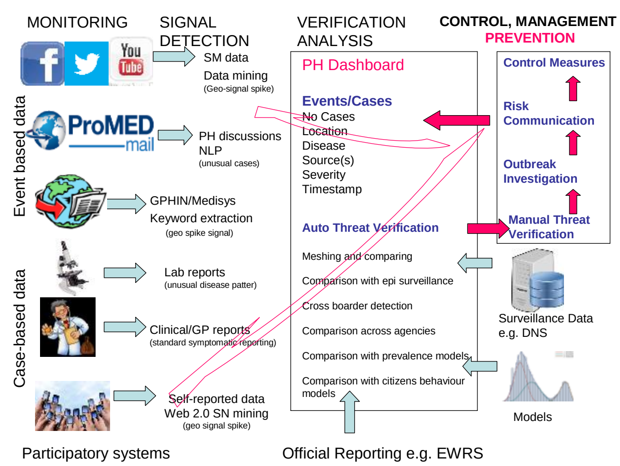

Figure 1. Integrated Digital Public Health

The infrastructure of our new *medi+board* system is illustrated in Figure 2.

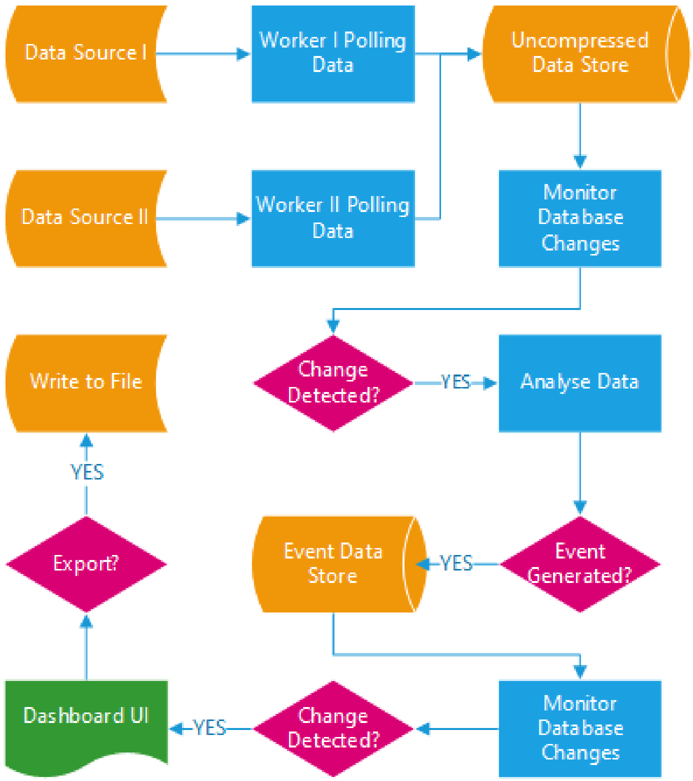

Figure 2. ***medi+board*** **system architecture**

This diagram shows the basic workflow of tracking data sources (as examples, Twitter, Medisys, HPA), scanning for signals, correlating with signals from other streams and generating events that are reported to the dashboard. As the framework is fully generic, the public health expert configures the system for use by defining the components using *event templates* that describe the logic of the components for tracking a particular disease according to available datasets and the ordering of pre-determined steps for risk assessment. These can include further investigation of data as well as organizing an emergency response or publicity. These *event templates* can be shared with users of equivalent permission once created, thus enabling the standardization of response procedure.

3.1 Signal Detection, Analysis and Correlation

Monitoring and detection of multiple channels requires different computational methods according to their structure and reliability – these are modularly provided (data mining, NLP, ML, data science, complex systems, social networks, etc). Newly identified signals in each data stream are validated by the system during which their *reliability coefficient* is adjusted. Once this source-specific processing has completed, results are cross-correlated with other data signals. During this process, reliability is adjusted again and events are formed that reflect situation awareness. This may even include cross-border detection as GPS-enabled streams could identify clusters of threats that would not traditionally trigger alerts in either region/country.

Our algorithm is expressed in the form of a *directed acyclic graph* which can be programmed by the analyst in a visual designer to prevent the need of writing code (Figure 3). We believe this approach makes understanding an algorithm provided by another analyst much easier as the graph itself directly visualizes the computational process. Apart from the resulting increase in productivity, every node encapsulates a specific step in the algorithm with its own distinct set of properties such as the *reliability coefficient* that can be manipulated directly in the designer. This makes tweaking settings more intuitive and illustrates how generic and customizable the algorithm is to aid in the solution to similar problems. Furthermore, we believe our graphs can be exploited to allow for novel approaches to debugging in future.

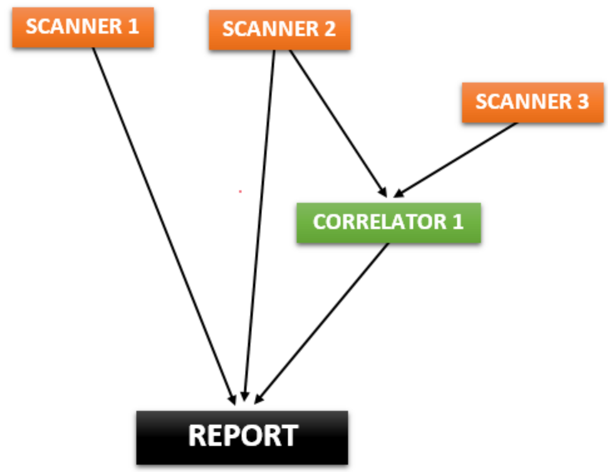

Figure 3. A simple example of an algorithm defined as a directed acyclic graph

3.2 Dashboard Visualization

Events verified by the processing system are forwarded to the dashboard and, according to the event template, presented in the appropriate visual format. Depending on the needs of the public health expert using the system, requests for further data or processing can be initialized as required.

Further dashboard segments include maps and other appropriate spatiotemporal visualization components, discussion forums for experts, reports and press releases which are all easy to navigate using an interactive timeline that forms the central organizational unit for data display. Figure 4 illustrates our design of the dashboard.

Figure 4. The *medi+board* dashboard screen

4. *medi+board* INFRACTRUCTURE

The overall infrastructure is illustrated below. It is designed to be fault-tolerant as all messaging is performed via scalable queues.

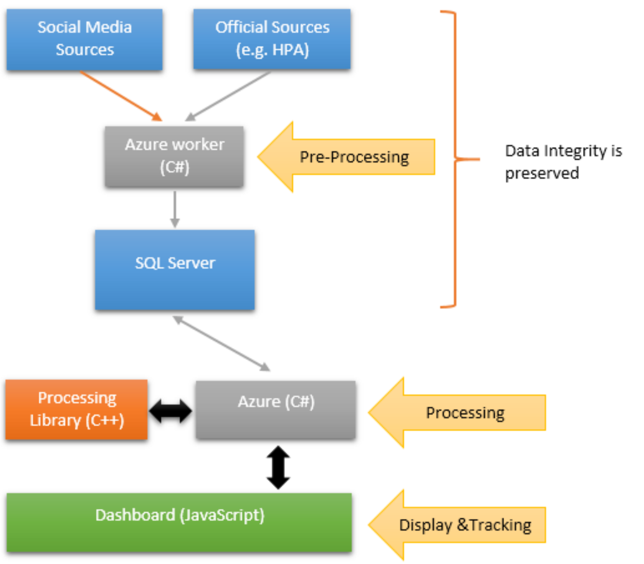

Figure 5. *medi+board* system architecture

As illustrated in Figure 5 the *medi+board* system is designed as a Windows Azure application. This enables it to scale relative to the amount of incoming data. Moreover, this scaling is dynamic and happens automatically, allowing us to leverage the full power of cloud computing.

We use the MapReduce paradigm to divide computation across nodes in order to achieve scalability while simultaneously improving reliability and fault-tolerance. Results are then aggregated in the dashboard. Mapping is achieved by having each individual node run an instance of our processing core that is responsible for detecting patterns such as potential correlations and formulating events accordingly. Flexibility and extensibility to add new diseases and monitor new trends is one of the key features of *medi+board* enabled by MapReduce and nodes can be programmed by the user by defining directed acyclic graphs. Each of these graphs can be evaluated independently as they are not allowed to reference each other. There *can* however be as many different instances as required, e.g. each looking for patterns pointing to a different kind of disease or symptom. Furthermore, other individual components of the system which are to be added in future (e.g., natural language processing) could simultaneously and independently work on incoming data items. The actual real-time processing and the dashboard represent the reduction step.

The dashboard itself is realized as an html5/JavaScript web application, meaning that it runs across different operating systems and does not require installation or local storage. One of the advantages of this model is that the public health professional is freed from the constraints of working in one fixed location.

Though the current iteration of the system is in the alpha-stages of its development, running simulations on data from the 2009 swine flu outbreak presented no computational problems for us.

5. SWINE FLU 2009 SIMULATION DEMO

In this section we present a demonstration of *medi+board* functionality on a simulation of the swine flu pandemics in 2009 made by running the system in desktop mode. We used three datasets: the HPA surveillance data from the Royal College of General Practice (RCGP), Google News API, and Twitter dataset collected during the pandemics in 2009 [12, 16, 22, 23].

During the simulation, potential threats are identified by scanning data from external files (rather than real-time sources), analyzing threats by a daily comparison of the number of cases against a *threshold coefficient* defined for each data source according to its reliability and a *cross-correlation coefficient* for each combination of sources to decrease the chance of false positives. Finally, analyzed and cross-correlated threats produce *events* that are updated on a daily basis as more information becomes available. The dashboard shows the three data streams in a dynamic way while statically illustrating other important segments of data, as if these were available, at two key points of the pandemics (containment phase in the UK when the demographics study of initial cases 'FF100' was conducted; and control phase in the autumn monitoring the distribution of the anti-virals as well as calls to the dedicated "fluline").

We show the dashboard displaying the results of this analysis in the "containment phase" in Figure 6. As explained in the previous section, the dashboard itself is dynamic, i.e. it is constantly updated as new data becomes available.

Figure 6. *medi+board* dashboard for the 2009 data set

6. FUTURE CHALLENGES

In future we plan to explore using additional data sources, surveillance datasets and signals identified by other public health agencies, as well as data from the Department of Health and the WHO. Furthermore, cross-validation of signals remains a challenge to which we hope our graph-based algorithms can make a significant contribution. The integration of accurate disease models and the spread of infection due to human travel into the validation algorithm is another key area of our future research.

Finally, it is worth noting that a large amount of data is not publicly available. There are two general reasons for this. Social media networks usually require explicit consent for making information visible to generic data mining algorithms (e.g. Facebook), thus, only public pages, such as brands, venues, agencies etc. could be tracked for EI purposes. Secondly, surveillance databases are legally 'country-owned' and require permissions for sharing even with the ECDC and the WHO (e.g. TESSY dataset). While the former problem is a fundamental issue of personal freedom and there are good reasons for keeping data private, the latter is predominantly political. We hope to make a strong case for collective epidemiological intelligence in order to combat diseases that pose a very real threat to the survival of the human race.

This means that sharing data in a machine readable format, in line with the Linked Data initiative in the UK (in the form of non-identifiable epidemiological datasets) at a national and international level is desirable. Furthermore, we strongly believe that given appropriate mechanism for the protection of privacy and personality, even company-internal datasets (such as mobile and pharmaceutics industry) could be exploited for research purposes.

7. CONCLUSION

Public health informatics is the driving force behind a paradigm shift in public health services. Realtime big data sources, citizen participatory systems and mobile digital traces generate a stream of location- and time-specific data to enhance traditional medical surveillance systems.

In this paper we introduced *medi+board* – a public health dashboard screening real-time data sources for early warning of infectious disease threats, cross-validating sources by correlating data streams and displaying results in an integrated format presented by means of an interactive dashboard. The system aims to significantly simplify the task of investigation and control of infectious diseases by public health experts.

Demonstrated on three data streams from the swine flu 2009 pandemic (RCGP surveillance in the UK, Google News and Twitter streams), the *medi+board* integrated public health dashboard provides a simulation illustrating how such system could substantially enhance future public health operations.

8. REFERENCES

[1] M Salathe, L. Bengtsson, TJ Bodnar, DD Brewer, JS Brownstein, C Buckee, EM Campbell, C Cattuto, S Khandelwal, PL Mabry, A Vespignani. Digital Epidemiology. PLoS Computational Biology 8(7): e1002616. dio:10.1371/journal.pcbi.1002616.

[2] SI Hay, DB George, CL Moyer, JS Brownstein. Big Data Opportunities for Global Infectious Disease surveillance. PLOS Medicine. April 2013, Volume 10, Issue 4, e1001413.

[3] P Kostkova. A roadmap to integrated digital public health surveillance: the vision and the challenges. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). 687-694., 2013.

[4] P Barboza, L Vaillant, A Mawudeku, NP Nelson, DM Hartley, LC Madoff, JP Linge, N Collier, JS Brownstein, R Yangarber, P Astagneu. Evaluation of Epidemic intelligence systems Integrated in the earl Alert in and Reporting Project for the Detection of A/H5N1 Influenza Events. PLOS One, March 2013, Volume 8, Issue 3, e57272.

[5] http://www.ot.co.uk/global/customers/successstories.htm?sys_action=show&id=661.

[6] WHO,http://www.who.int/csr/alertresponse/epidemicintelligence/en/index.html.

[7] Linge JP, Steinberger R, Weber TP, Yangarber R, van der Goot E, Al Khudhairy DH, Stilianakis NI. Internet surveillance systems for early alerting of health threats. EuroSurveill. 2009;14(13):pii=1916. (2009).

[8] http://healthmap.org/en/?gclid=COmHjfHtiKwCFdQLfAodcH-V-Q.

[9] M. Keller, M. Blench, H. Tolentino, C.C. Freifeld, K.D, Mandl, A. Mawudeku, G. Eusenbach, J.S. Brownstein. Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. Emerging Infectious Diseases, Vol. 15, No. 5, May 2009, doi: 10.3201/eid1505.081114.

[10] Google Flu Trends, http://www.google.org/flutrends/

[11] J. Ginsberg, M. H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant. Detecting influenza epidemics using search engine query data. Nature Vol 457, 19 February 2009, doi: 10.1038/nature07634.

[12] S. Wiseman, P. Kostkova, E. de Quincey, G. Jawaheer. Providing guidance during the swine flu outbreak in 2009: An evaluation study of the National Resource for Infection Control (NRIC). Poster and Abstract in the Proceedings of the 14th International Conference on Infectious Diseases (ICID), March 9-12th 2010, Miami, Florida, US.

[13] P Kostkova, D Fowler, S Wiseman, JR Weinberg, Major infection events over 5 years: how is media coverage influencing online information needs of health care professionals and the public? *J Med Internet Res 2013;15(7):e107)* doi:10.2196/jmir.2146.

[14] Lamos, V., de Bie, T., and Cristianini, N. "Flu Detector – Tracking Epidemics on Twitter". In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010), pp. 599-602.

[15] Lampos, V. and Cristianini, N. "Tracking the flu pandemic by monitoring the Social Web". In Proceedings of the 2nd IAPR Workshop on Cognitive information Processing (CIP2010), pp 411—416, 2010.

[16] Szomszor, M., Kostkova, P., de Quincey, E. (2010). *#swineflu: Twitter Predicts Swine Flu Outbreak in 2009*. M Szomszor, P Kostkova (Eds.): ehealth 2010, Springer Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering LNICST 69, pages 18-26, 2011.

[17] V Lampos, N Cristiani. Nowcasting Events from the Social Web with Statistical Learning. ACM TISM, Vol 3, No 4, Article 72, September 2012.

[18] A Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. ACM, Proceedings of the SOMA '10, pages 115-122.

[19] J Gomide, A Veloso, W Meira Jr, V Almeida, F Benevenuto, F Ferraz, M Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, ACM, Web Science 2011, June 14-17, 2011, Germany.

[20] M Salathe, C.C. Freifeld, S R Mekaru, A.F Tomasulo, J.S. Brownstein. Influenza A (H7N9) and the Importance of Digital Epidemiology. The New England Journal of Medicine. July 3rd 2013, nwjm.org.

[21] A Signorini, AM Segre, P. M. Polgreen. The Use of Twitter to track levels of Disease Activity and public Health Concern in the U.S. during the Influenza a H1N1 Pandemic. PLOS One. 2011; 6(5): e19467, doi:10.1371/journal.pone.0019467.

[22] E de Quincey, P Kostkova Early warning and outbreak detection using social networking websites: The potential of twitter, 2nd ehealth conference, Istanbul, Turkey, LNICST Electronic Healthcare, p21-24, 2009.

[23] M Szomszor, P Kostkova, C St Louis. Twitter informatics: Tracking and understanding public reaction during the 2009 swine flu pandemic. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011, pp. 320 – 323.