Social Web Information Monitoring for Health

June 26th, 2009

Alessio Signorini
alessio-signorini@uiowa.edu

The Internet is Growing

In October 2008 more than 82% of U.S. households had a computer and 92% of them used it to access the Internet.

People spend on average 48 hours per month online, reading news (46%), paying bills (40%), shopping (37%), booking flight tickets (20%), …

Many users actively participate to online communities: 75% have profiles on Facebook or Myspace, 15% uses blogs or forums, 30 Million use Twitter, and many more consult Wikipedia.

The Web seen as a Giant Library

The growth of the Internet also increased the amount of information accessible to the general public on any topic.

Historical data, maps, graphs, and many other resources are available online for free. Many Encyclopedias and other publications exist today only in electronic form.

More than 20% of Americans look for medical advices online. Health domains (e.g., WebMD, MayoClinic, …) are among the most popular sites of the Internet, together with medical support groups.

Lots of User Data are Available

Many of today's activities are performed electronically: reading news, paying bills, shopping with credit cards, email, booking vacations, …

The recent decrease in costs of disk and cpu power coupled with the expansion of cloud computing, allow for all these transactions to be saved in digital form.

Data Mining techniques are applied on aggregated data to extract patterns, common trends and detect anomalies. Credit card companies, insurances and online retailers heavily rely on log data analysis.

What is the Social Web?

During its first years, the Internet was considered a library with lots of consumers and just a few publishers.

In the last years the Internet became more socially active (“Web 2.0”). Today, everyone is at the same time a consumer and publisher of data.

Every day, millions of people update their status on Facebook, upload some new pictures on Flickr, send a couple of Tweets, listen to songs on Pandora, share some news on Digg, write a blog post or comment on it.

The Social Web, May 2009

18.8 Billion

Minutes spent on Facebook/MySpace

9.4 Billion

Searches performed

27.2 Million

Blog Posts

380.5 Million

Twitter Messages

Goal of this Research

This project focus on the health sciences, collecting, studying and validating available data as an additional signal to monitor and better manage disease outbreaks.

- Query Logs

  Correlate people's searches with health conditions
- Blog Posts

  Identify perception of general public on health topics
- Social Status Updates

  Monitor the real-time, geo-located stream for health clues
- Proxy Logs

  Observe variation in traffic on health websites' pages

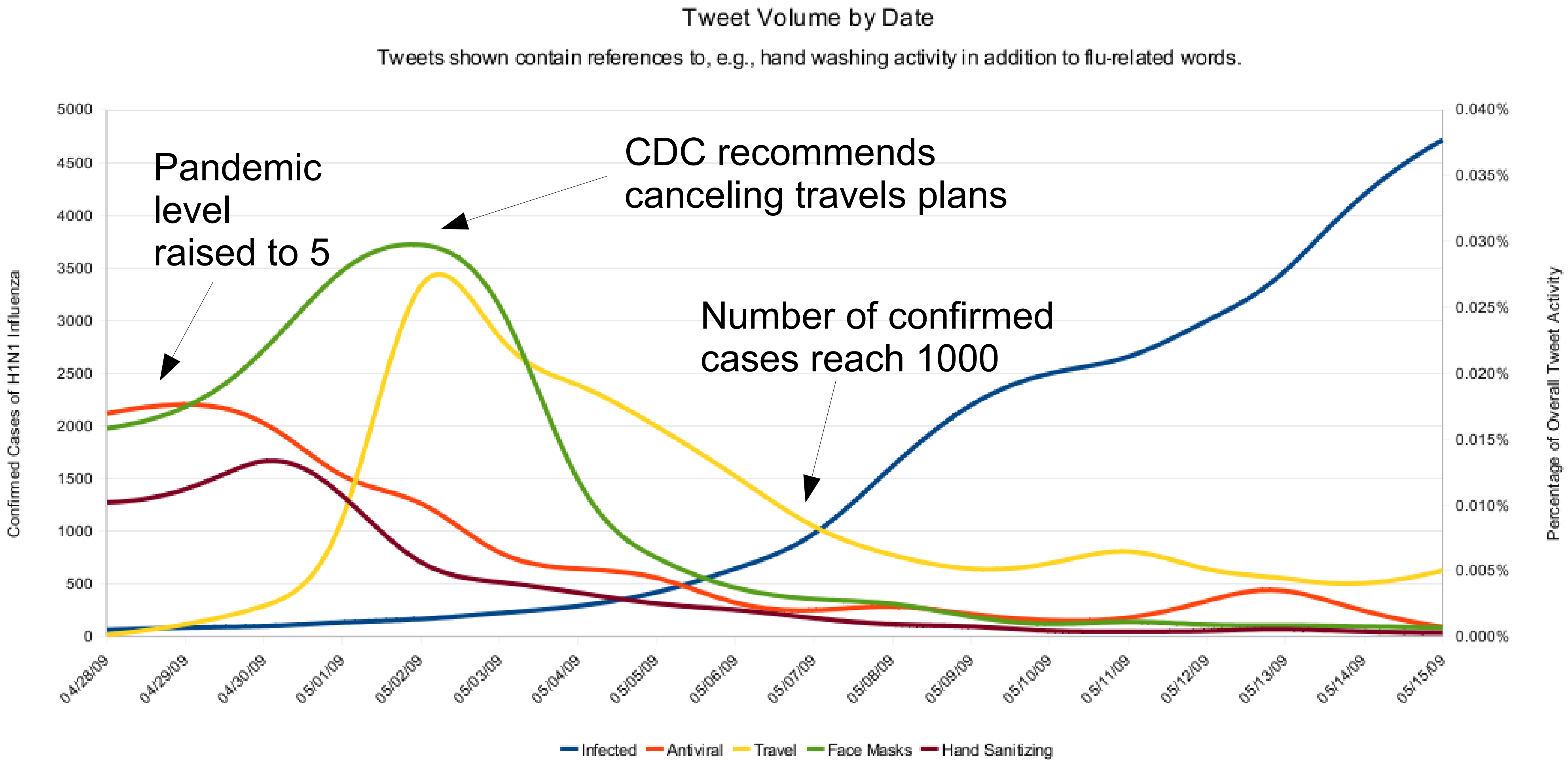
Results: H1N1 Outbreak

In April 2009 a new flu virus of swine origin has been detected in Mexico. The virus is capable of infecting human and the illness sparked all over the United States.

The CDC promptly responded to the outbreak releasing frequent updates, reports on antivirals use, and general suggestions to reduce infection risks.

The public response has been massive. Thousands of news articles were published every day, travels were canceled, and people stockpiled on food and antivirals.

Results: H1N1 Outbreak



Tweet Volume by Date

Tweets shown contain references to, e.g., hand washing activity in addition to flu-related words.

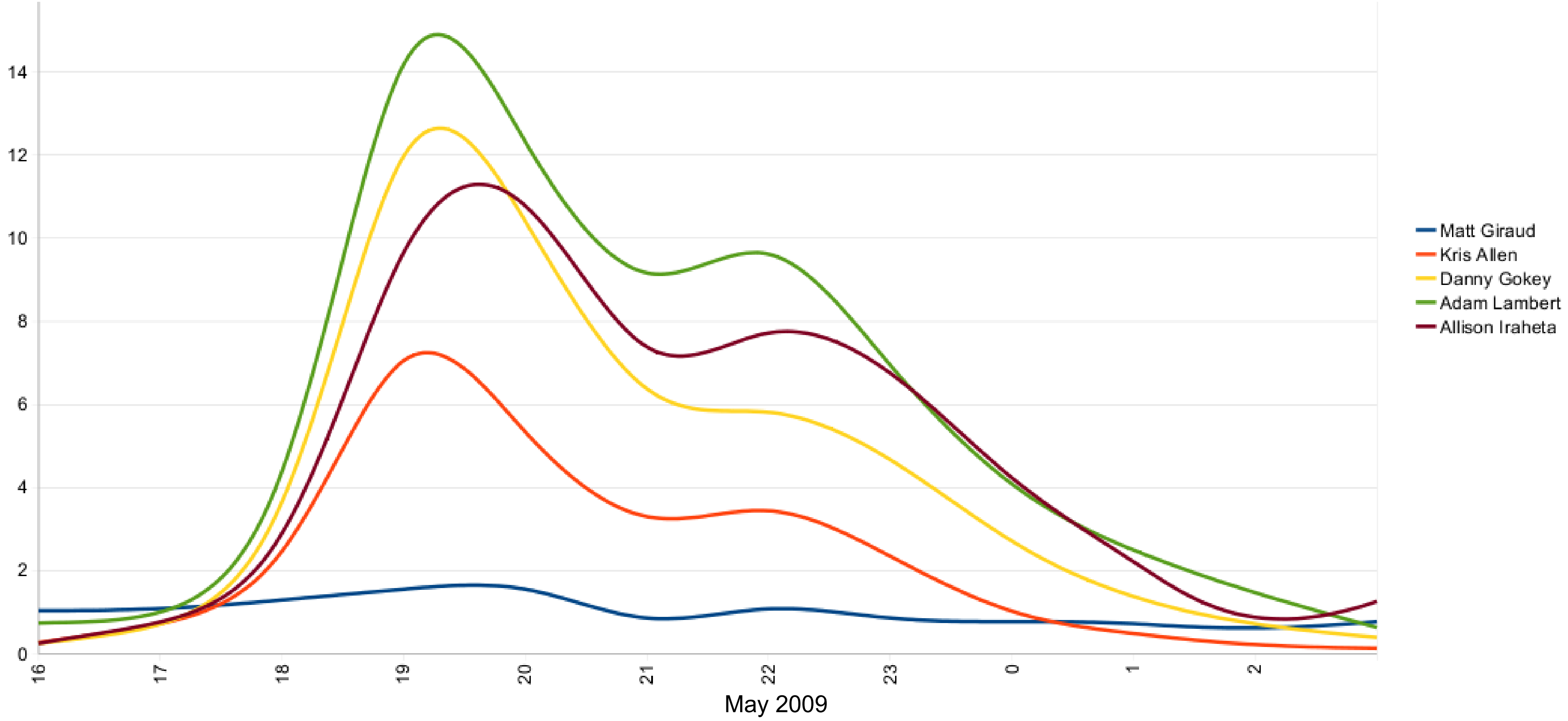
Pandemic level raised to 5

CDC recommends canceling travels plans

Number of confirmed cases reach 1000

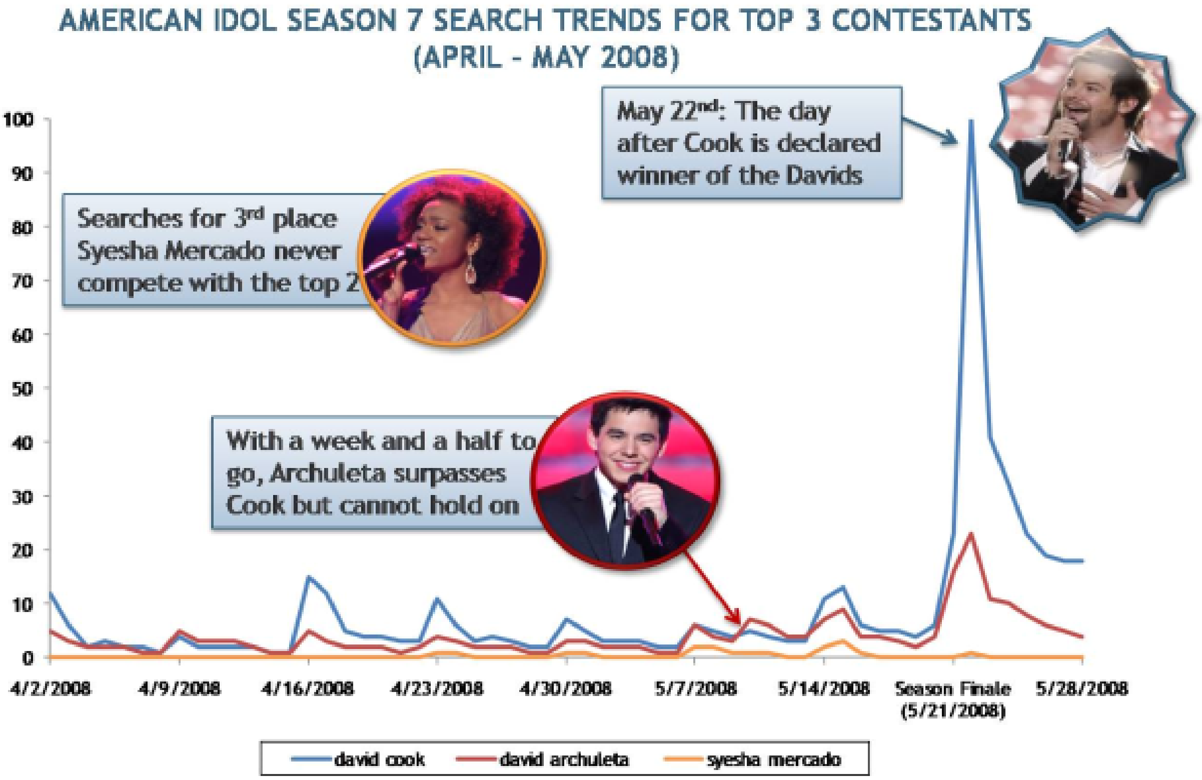
Results: American Idol 2009

American Idol is a reality-show competition to find new solo musical talent very popular in the United States. The target audience is similar to the users of social networks.



Results: American Idol 2009

The tech-blog Mashable predicted that Adam Lambert would win the competition analyzing and comparing Google Trends graphs of the singers' names.



AMERICAN IDOL SEASON 7 SEARCH TRENDS FOR TOP 3 CONTESTANTS (APRIL - MAY 2008)

May 22nd: The day after Cook is declared winner of the Davids

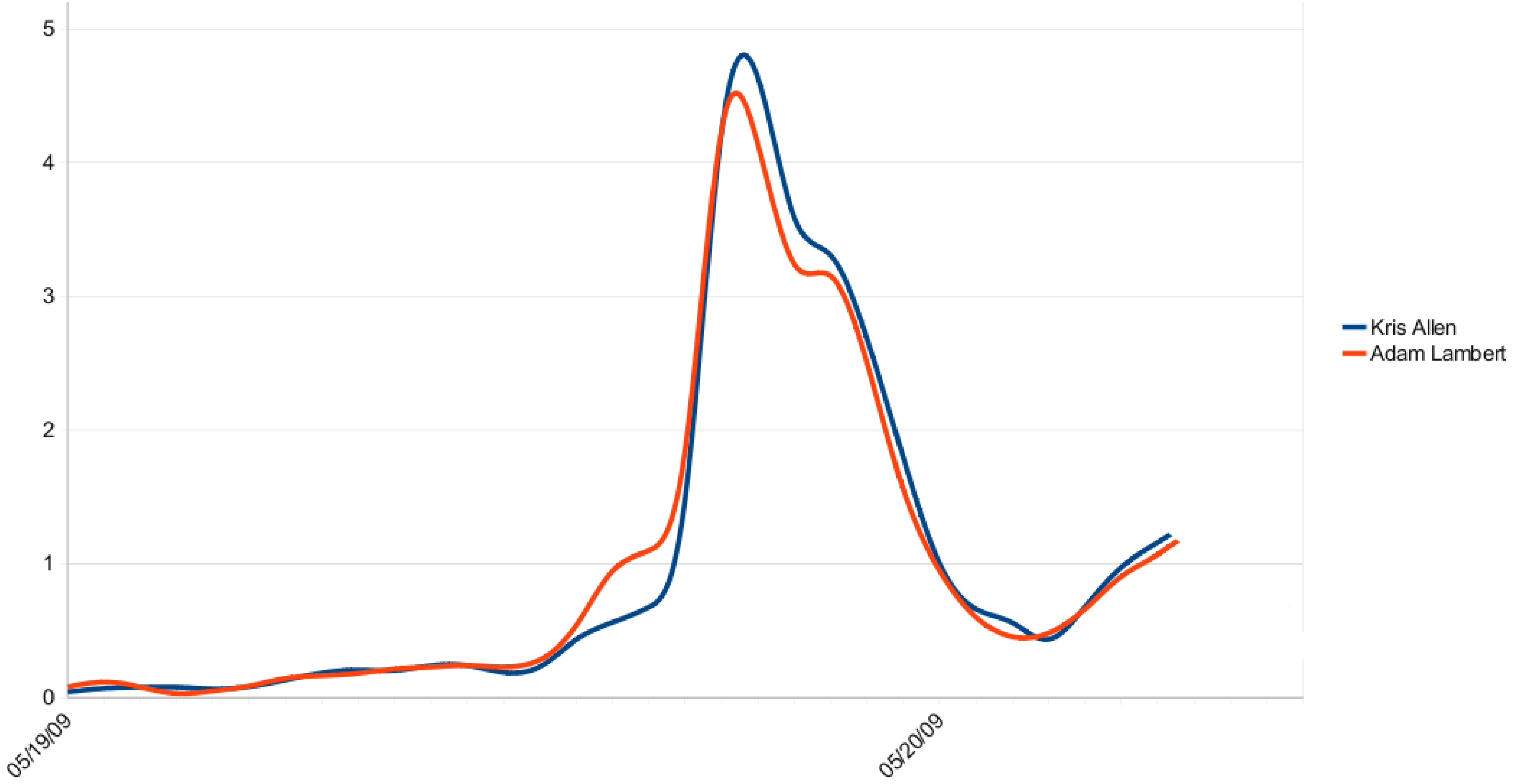
Searches for 3rd place Syesha Mercado never compete with the top 2

With a week and a half to go, Archuleta surpasses Cook but cannot hold on

Source: 360i Analysis Google Insights; May 2009

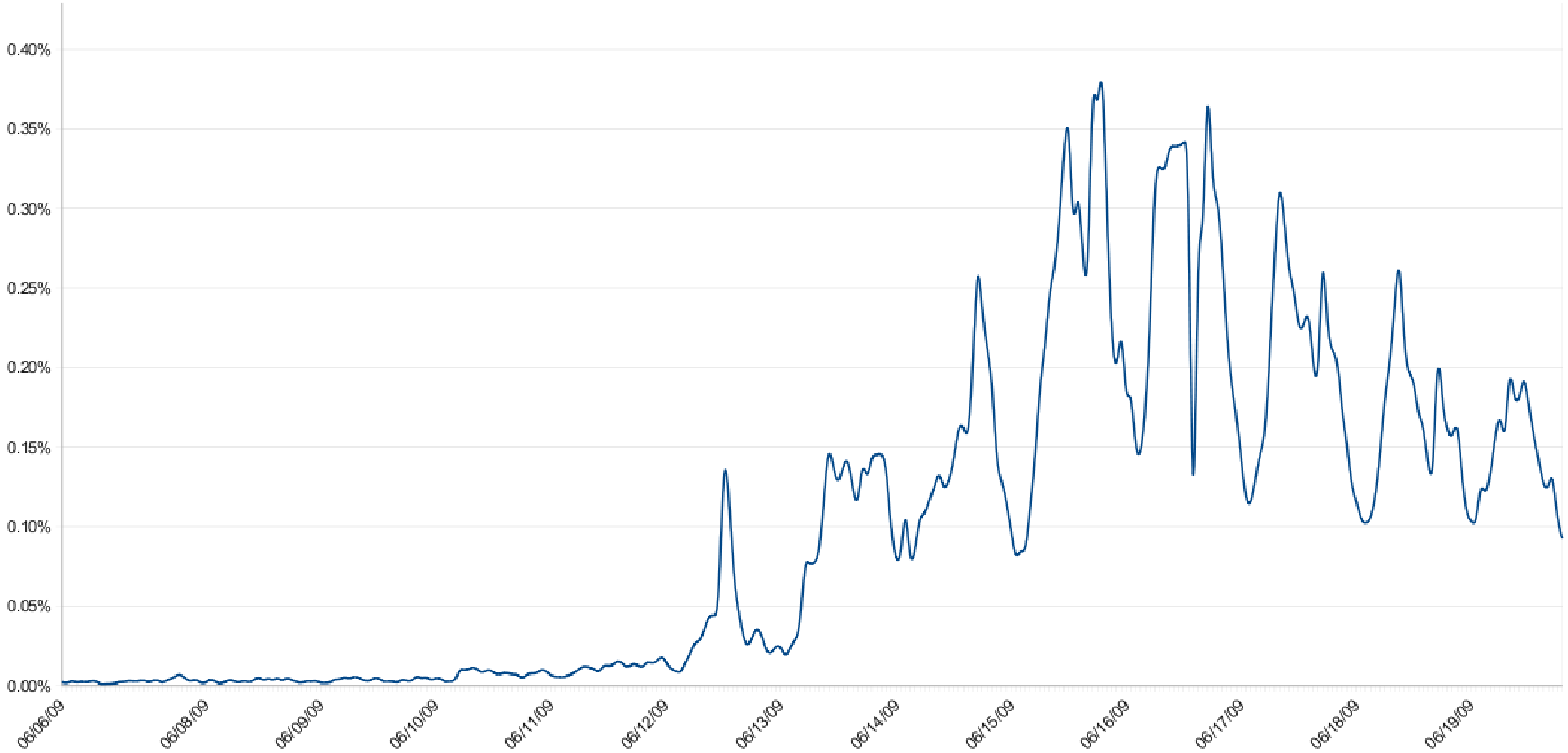
Results: American Idol 2009

We compared the relative number of positive tweets received right after the season final's performance for the two contestants and declared Kris Allen the winner.



Results: Iranian Elections

On June 12th, 2009 was held the tenth Iranian Presidential election. The European Union expressed concerns on irregularities and insurrections sparked on the streets.



Future Directions

The results obtained are very encouraging. Social web activities seems highly correlated with public perception of certain topics and the information implicitly released could be very effective in a public health context.

- Twitter and Blog Posts

  Introduce classification and consider geo-location

- Wikipedia

  Monitor page traffic and number of edits

- Browsing History and Search Queries

  Use query classification and analyze page traffic variations

Future Directions

We will analyze many sources of social activity data (e.g., Twitter, Wikipedia, ...) in search of trends and patterns which can be directly correlated to the public sentiment.

Well-known data mining techniques (e.g., classification, clustering and entity extraction) will be used to extract and identify the features and characteristics of each trend.

Our experiments will be correlated to current events (e.g., seasonal flu trends) and validated against official health data (e.g., CDC flu reports) so that any findings can be directly applied by health authorities.