SAND2017-7838C

Temporal Anomaly Detection in Social Media

Jacek Skryzalin, Richard Field, Andrew Fisher, Travis Bauer

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Social Media…

Day	Top trends
Monday	#mondaymotivation, #blackoutday, #NationalOreoCookieDay, #SXSWEdu, #ARMYSelcaDay
Tuesday	#Vault7, #NationalPancakeDay, #Trumpcare, Tom Price, #TuesdayMotivation, #WhileWaitingForYourTextBack
Wednesday	#InternationalWomensDay, #GoogleNext17, #SheInspiresMe, #EmbarrassedToAdmitIveNever, #wednesdaywisdom
Thursday	#RIPBIG, #ThursdayThoughts, #NationalMeatballDay, #WeirdThingsToCompliment, Torrey Smith
Friday	#buffyslays20, #SXSW, #FridayFeeling, #MakeAFilmUpbeat, Purdue, #FlashbackFriday

Social Media…

I like #carrots

We like #carrots

Textual analysis of social media

Instance-based approaches:

- Physical (velocity)
- Statistical (chi-squared)
- Automaton (meme-tracker)

Bayesian approaches:

- Topics over time
- Dynamic topic models
- Online LDA

Textual analysis of social media

Instance-based approaches:

- Physical (velocity)
- Statistical (chi-squared)
- Automaton (meme-tracker)

Bayesian approaches:

- Topics over time
- Dynamic topic models
- Online LDA

Given timestamped documents, what ***trends or topics*** characterize each time interval?

Textual analysis of social media

Instance-based approaches:

- Physical (velocity)
- Statistical (chi-squared)
- Automaton (meme-tracker)

Bayesian approaches:

- Topics over time
- Dynamic topic models
- Online LDA

Given timestamped documents, what ***trends or topics*** characterize each time interval?

Given timestamped documents, what can we discover about the ***time intervals*** from latent trends or topics?

- Even if our sampling is not entirely reliable?

Our approach – PAKL

- Study information theoretic differences between current term distributions and the baseline term distribution.
- There's always *something* trending, but more significant trends will cause a greater divergence from baseline.

Our approach – PAKL

- Assume that the baseline term distribution is q, and that the current term distribution is p. The *Kullback-Leibler* divergence is defined as

$$KL(p||q) = \sum_{w} p(w) \log \frac{p(w)}{q(w)}$$

- Each *summand* of $KL(p||q)$ measures how much information w carries relative to baseline.

Our approach – PAKL

- Assume that the baseline term distribution is q, and that the current term distribution is p. The *Kullback-Leibler* divergence is defined as

$$KL(p||q) = \sum_{w} p(w) \log \frac{p(w)}{q(w)}$$

- Problem: for either $p(w) \approx q(w)$ or $p(w) \approx 0$, the summand corresponding to w is approximately 0.
- We'd like to capture both increases *and* decreases in usages of terms.

Our approach – PAKL

- Define a pointwise antisymmetric Kullback-Leibler score via:

$$PAKL_w(p||q) = \big(p(w) + q(w)\big) \ln\left(\frac{p(w)}{q(w)}\right)$$

Our approach – PAKL

- Define a pointwise antisymmetric Kullback-Leibler score via:

$$PAKL_w(p||q) = \big(p(w) + q(w)\big) \ln\left(\frac{p(w)}{q(w)}\right)$$

- Words for which $PAKL_w(p||q)$ is very *positive* are being used *more* frequently in p than in q.
- Words for which $PAKL_w(p||q)$ is very *negative* are being used *less* frequently in p than in q.

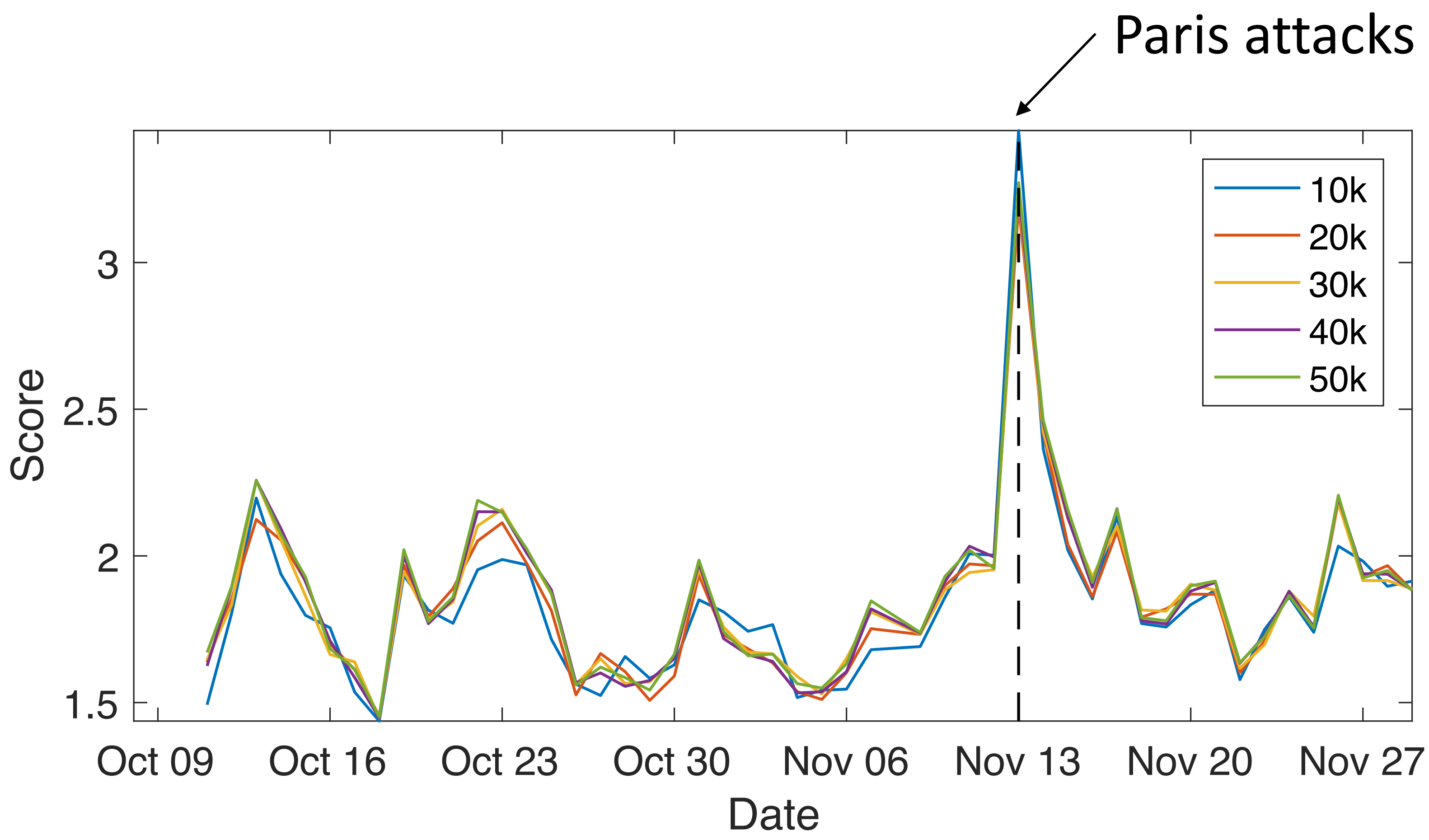
Our approach – PAKL

- Define a pointwise antisymmetric Kullback-Leibler score via:

$$PAKL_w(p||q) = \big(p(w) + q(w)\big)\ln\left(\frac{p(w)}{q(w)}\right)$$

- Words for which $PAKL_w(p||q)$ is very *positive* are being used *more* frequently in p than in q.
- Words for which $PAKL_w(p||q)$ is very *negative* are being used *less* frequently in p than in q.
- There is signal in the sum of the *n* highest PAKL scores for each time period.

PAKL scores are robust to size of dataset.



Extraction of important documents

Important terms (articles and prepositions removed):

Nov 13: paris, #prayforparis, #madeintheam, prayers, attacks

Nov. 26: Thanksgiving, thankful, happy, #mtvstars, britney

Important documents:

Nov 13: Sending prayers to the people in Paris #PrayForParis

Nov. 26: thankful for everything <emoji> Happy Thanksgiving !!

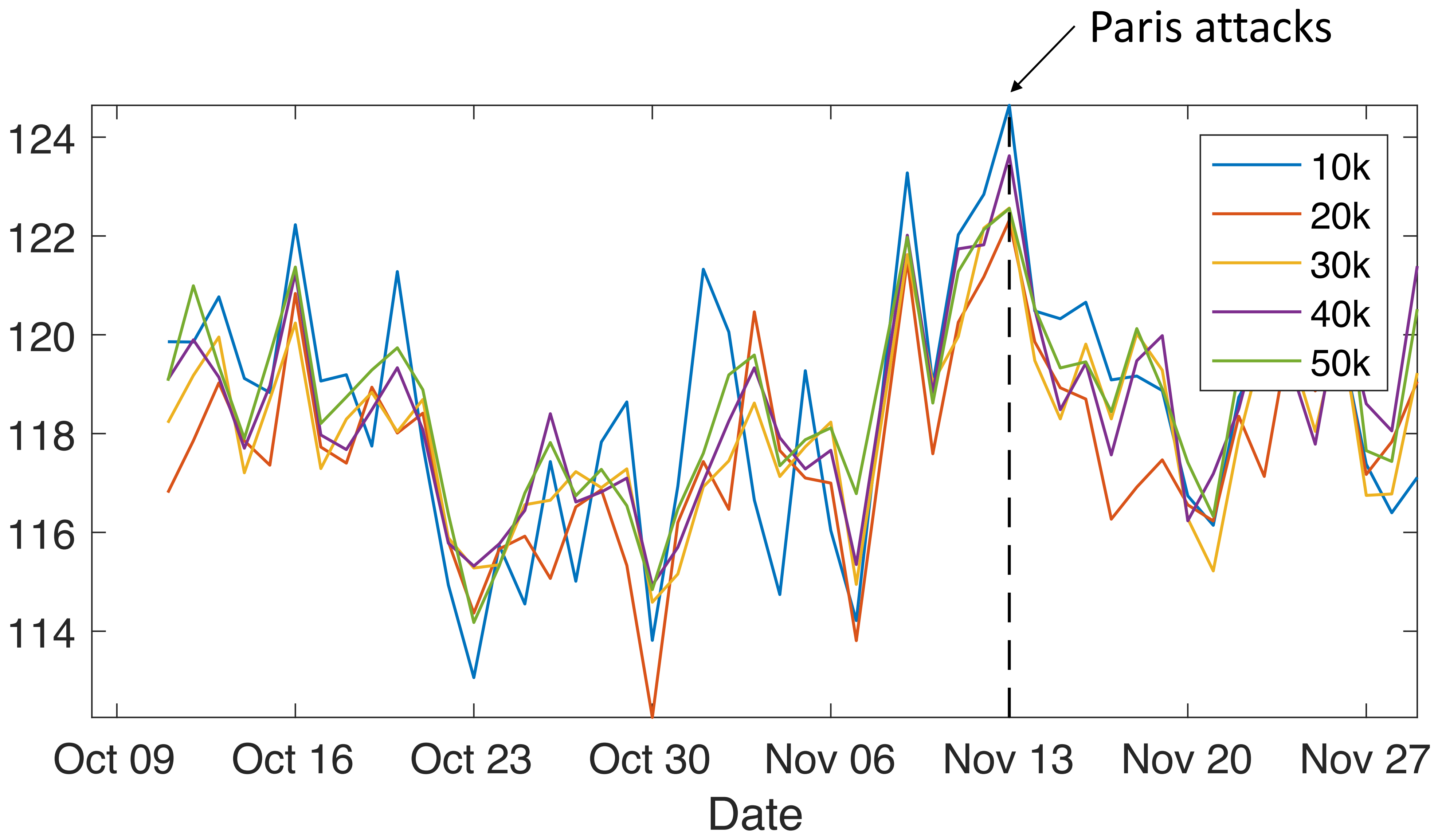
Our approach – Cluster coherence

Algorithm:

- For each document, create a vector by taking a tf-idf-weighted average of term vectors (e.g., GloVe, word2vec).
- Perform spherical clustering on the resulting document vectors.
- Measure cluster coherence.

Higher cluster coherence indicates that the topics being discussed are more tightly focused, indicating heightened state.

Cluster scores are robust to size of dataset



Probabilistic Feature Fusion

It is best to fuse the scores produced by each weak indicator in order to create a more robust, more accurate system.

- To fuse scores s_i, generated during time period t, with weight w_i into a final (fused) score, we compute:

$$\Gamma = \sum_i w_i \log\big(1 - F_i(s_i)\big)$$

Probabilistic Feature Fusion

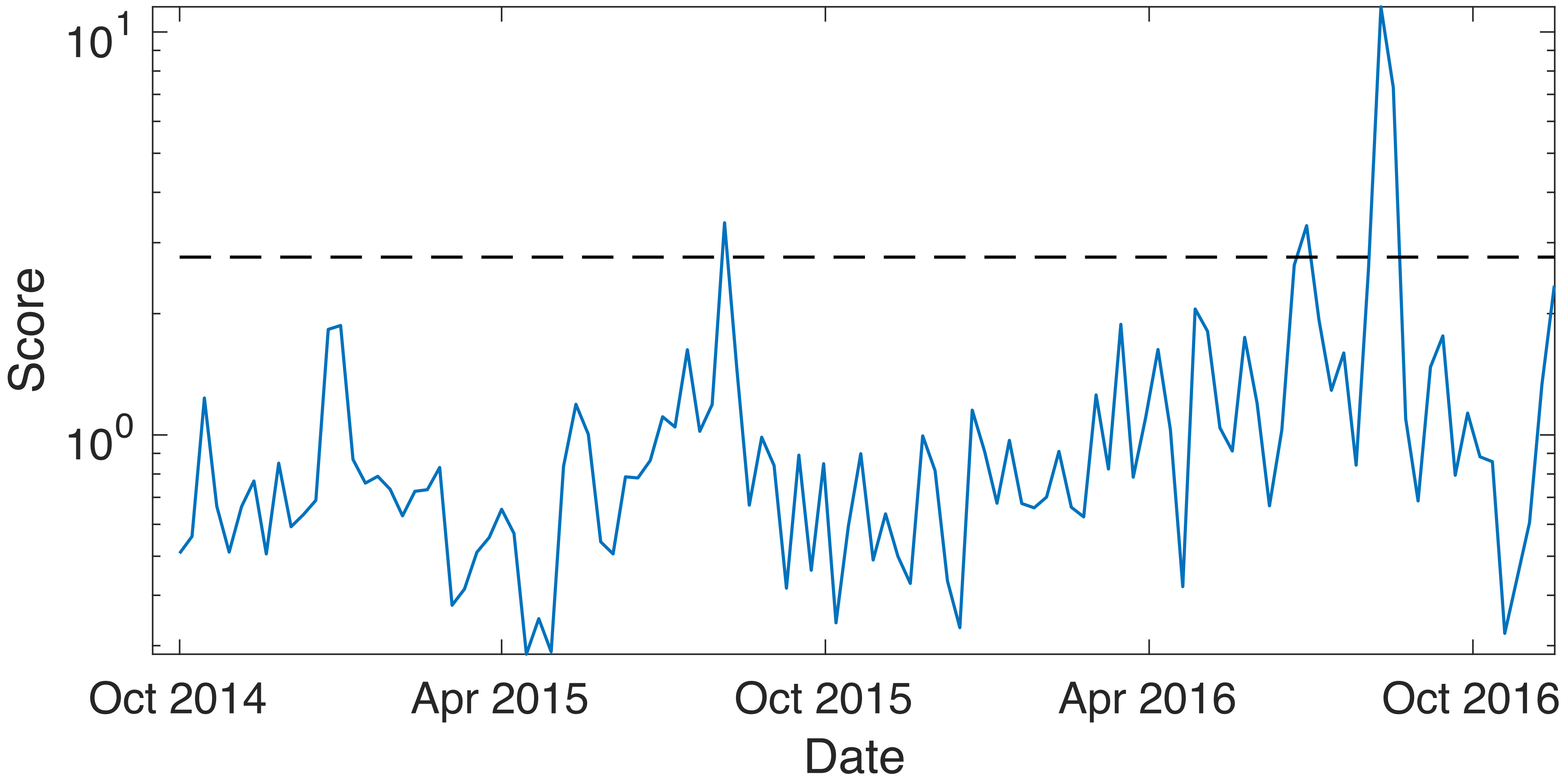
It is best to fuse the scores produced by each weak indicator in order to create a more robust, more accurate system.

- To fuse scores s_i, generated during time period t, with weight w_i into a final (fused) score, we compute:

$$\Gamma = \sum_i w_i \log\left(1 - F_i(s_i)\right)$$

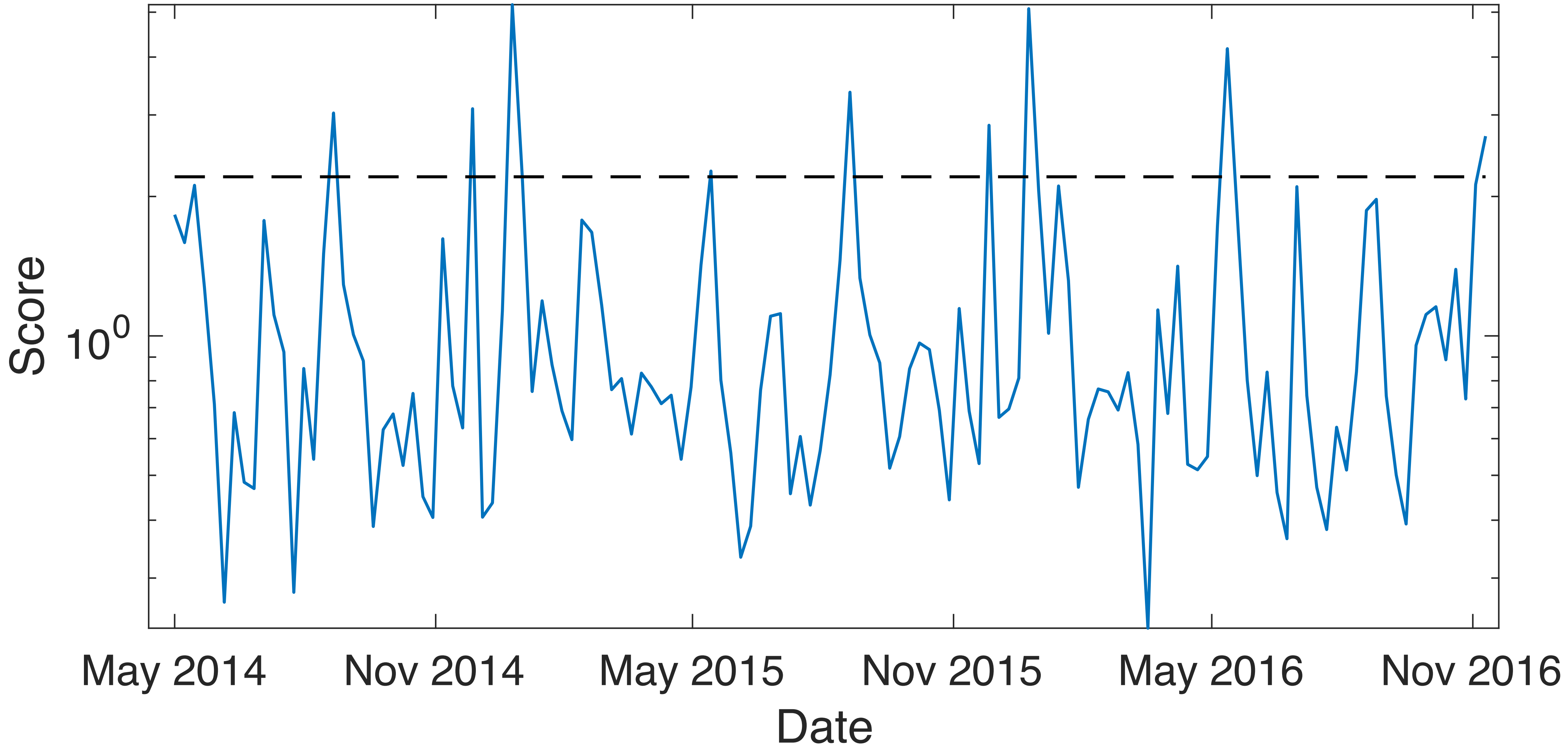
Γ is modeled as a gamma distribution whose parameters can be calculated. This underlying gamma distribution can then be used to assess significance.

Fused Scores (Olympics)



Dataset: a collection of tweets from Olympians and Olympics professionals.

Fused Scores (Universities)



Dataset: a collection of tweets from US universities.