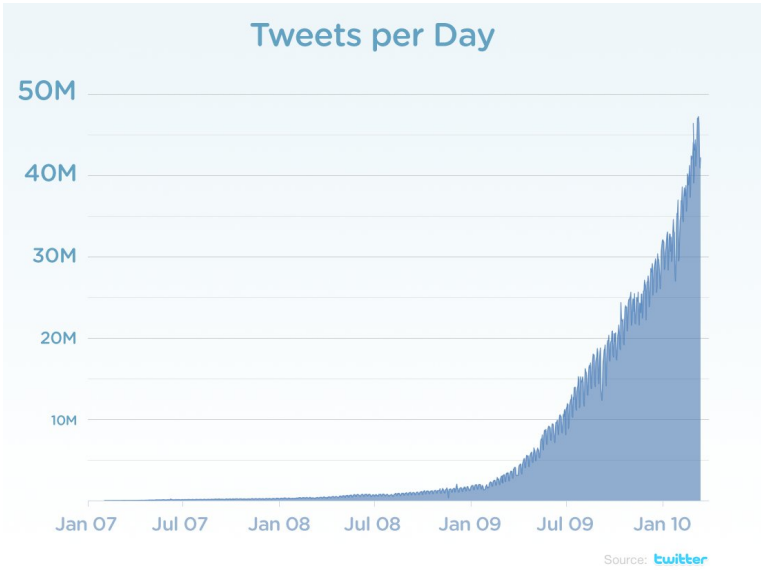
Inferring User Interests from Tweet Times

Dinesh Ramasamy, Sriram Venkateswaran and Upamanyu Madhow

Department of ECE, University of California Santa Barbara

Lots of data



More and more *users*

More and more *user generated* data

Detailed understanding

- Interests of ***each*** user
  - "*Whose feed* should I place my advertisement on?"
  - "*Which post* should come up first on a users' feed?" (user experience)
- Value to ***simple methods***
  - Used ***repetitively*** over a large user pool

Existing methods

Obtain user interests via:

- **Analysis of user generated content**
  - Tweets/posts
- User lists/groups
- Structure of user-interactions in the network
- This work
  - Time → Infer user interests

Select references

- TwitterStand [Sankaranarayanan et al]
  - Cluster tweets into different news groups
- Locate earthquakes in space & time [Sakaki et al]
  - *When* earthquakes happened?
- PET [Lin et al]
  - Identify trending topics, *when* trending?

Live conversations

- Always connected
  - Smartphones & tablets

- Broadcast current thoughts
  - Live commentary on “current events” of interest

- Prompts *live conversations* with others who share similar interests

Value to synchrony

Information in time

- Twitter amplifies *live*
  - Brief nature of tweets

Tweets times tied to "current events"

- Identify interests from time of tweets
  - Knowledge of times of "events"
  - Users whose tweet times "correlate" with these times

Example

- Tweet times tied to external events

"Dumb and Dumber To"
Filming started on Sep 24th

- Use *known times of external events* to learn user interests

Interests from tweet times

Sep 24th : Filming starts; Movie set photos released

Movie promo release?

?

?

t

Sep 1st : Casting decisions

I am so stoked about the new Dumb and Dumber movie!

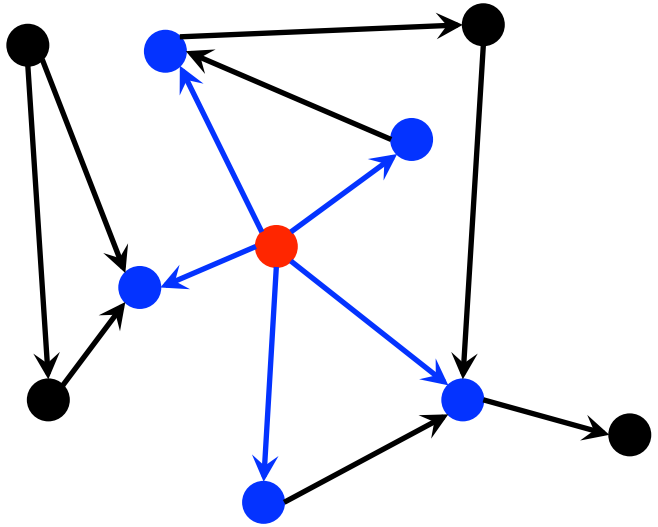
Reply Retweet Favorite More

1 FAVORITE

1:50 PM - 24 Sep 13 from

User interactions

- People you interact with share *some* of your interests



- Clues from ***tweet times*** of neighboring users

Overview

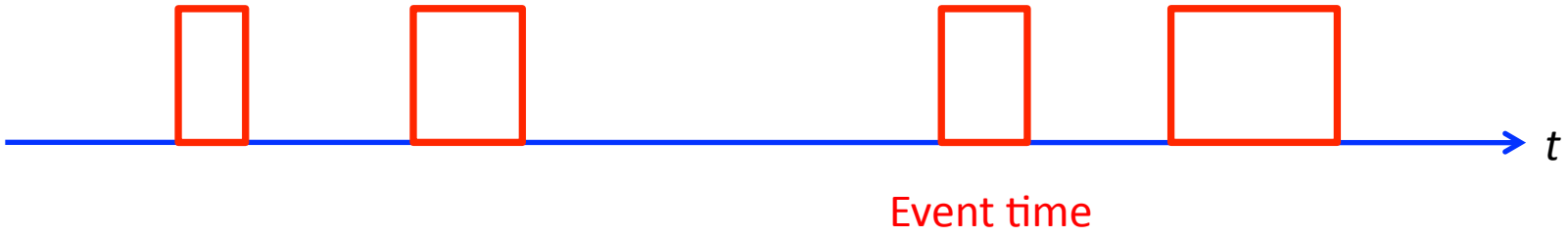
- Interests from tweet times of user
- Incorporate tweet times of neighbors
- Limitations
- Future work

INTERESTS FROM TWEET TIMES

Interests from tweet times

Buzz about topic X at certain times

– "Event times" *known* to us



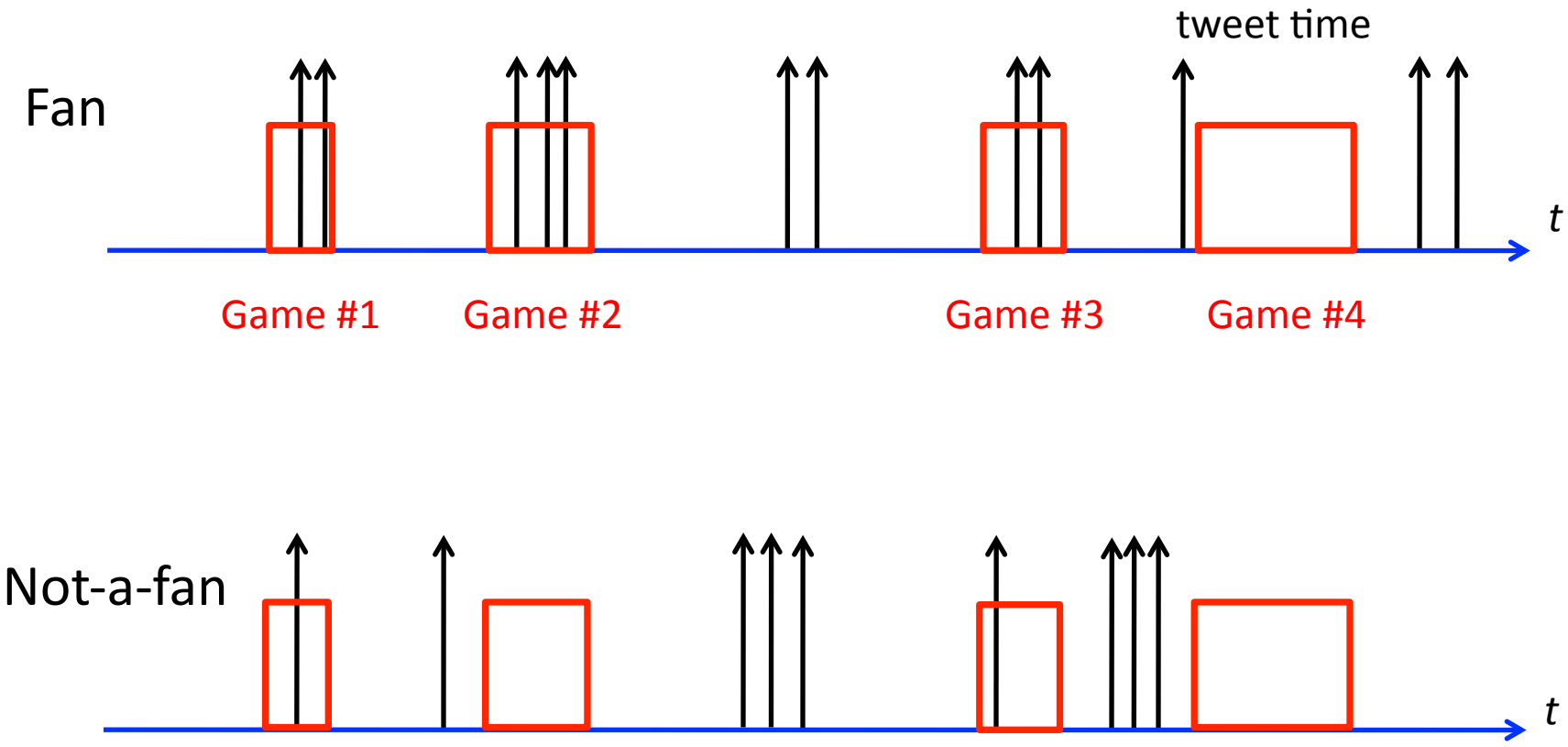
Expect users interested in X to tweet during these event times

Baseball fandom

Want to identify fans of a baseball team

Game times ←→ event times

Timeline of tweets



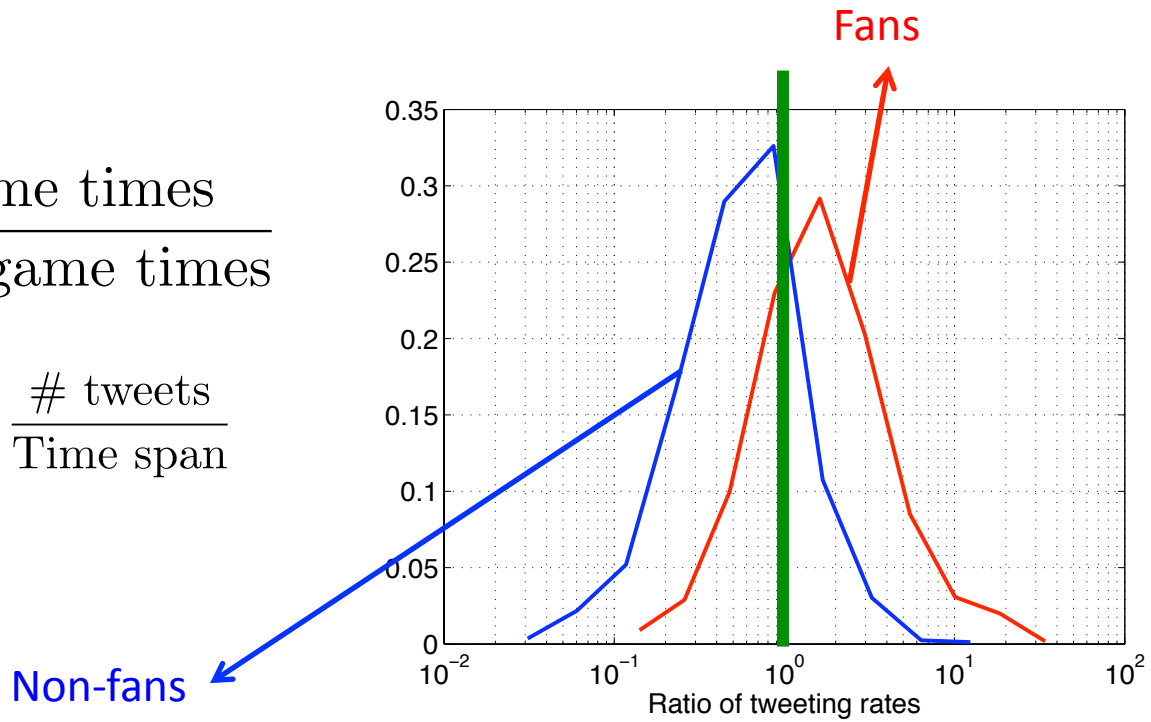
A fan tweets **more often** during game times

More often than?

- Some users tweet prolifically; others hardly tweet
  - Need a **personal "baseline"**
- More often when compared to other times for the *same user*

$$\frac{\text{Rate during game times}}{\text{Rate during non-game times}}$$

$$\text{Rate} = \frac{\#\text{ tweets}}{\text{Time span}}$$

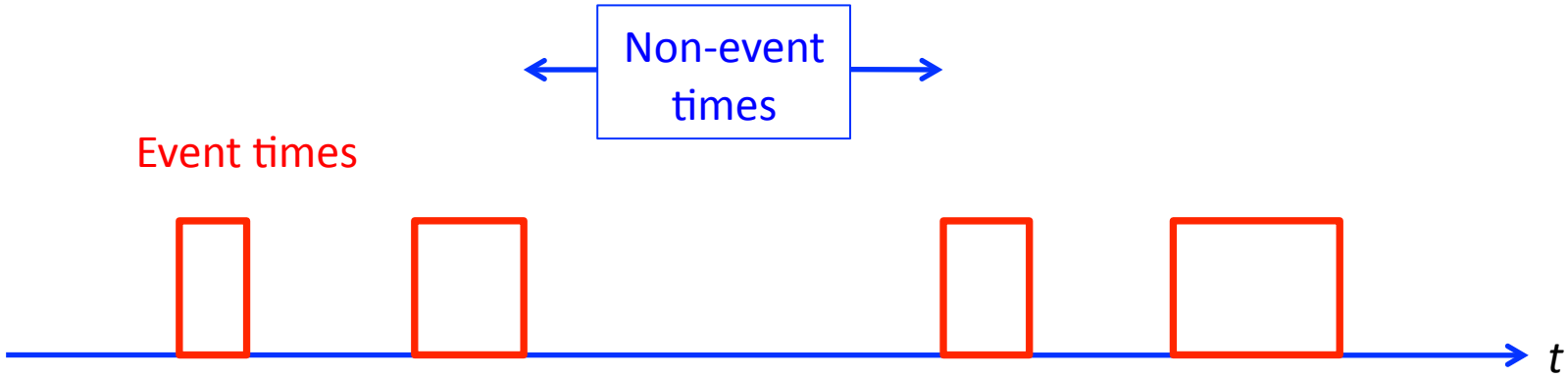


Ratio of rates

- "Ratio of rates" statistic $\frac{\text{Rate during game times}}{\text{Rate during non-game times}}$

- Prolific user
  - # tweets during games : 12 (30 hours)
  - # tweets during non-games times : 24 (150 hours)
- Sporadic user
  - # tweets during games : 1 (30 hours of games)
  - # tweets during non-games times : 2 (150 hours)
- Same ratio of rates – 5/2
  - Intuition: **More confident about the prolific user**

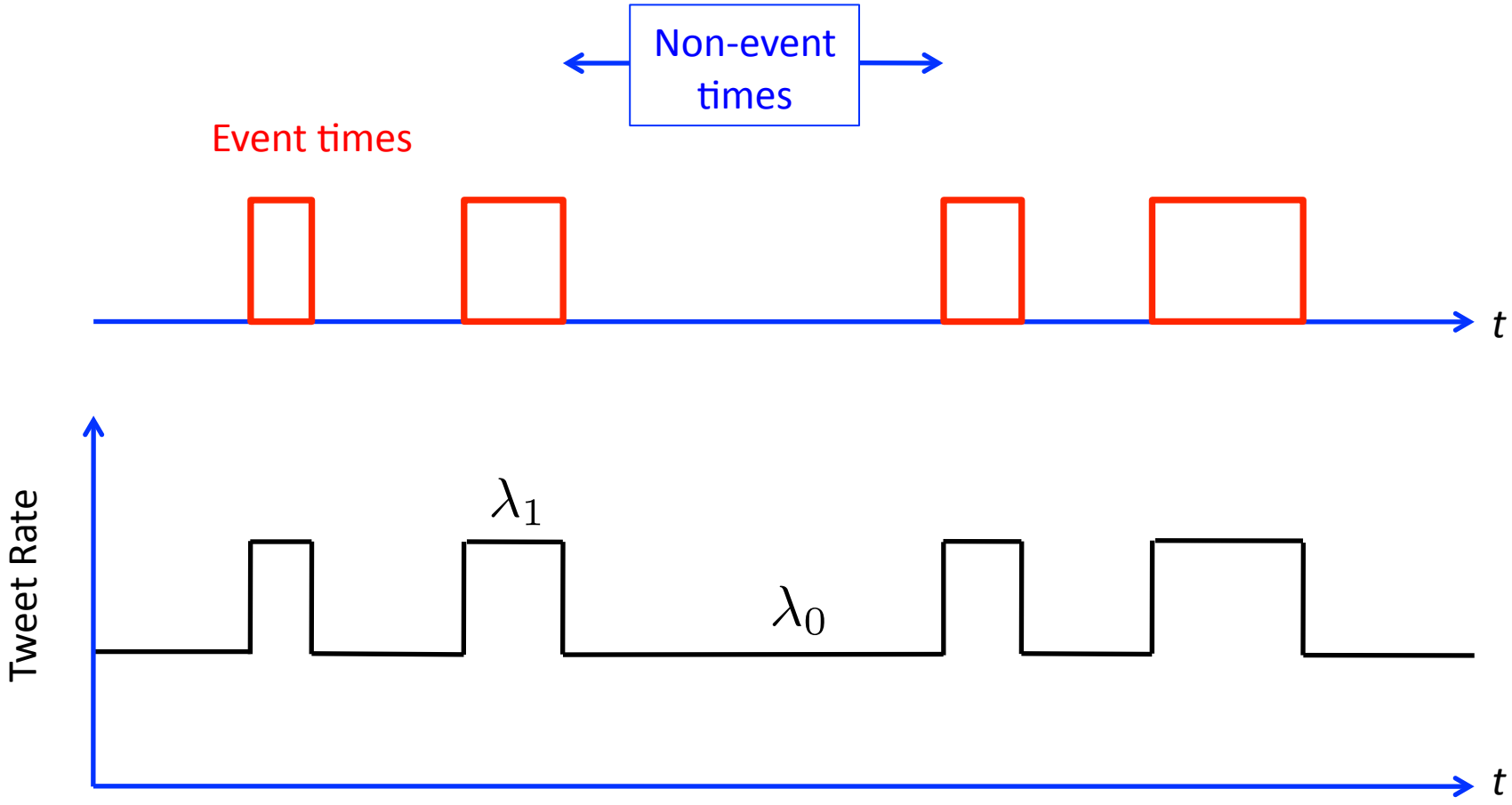
Statistical model for tweet times



Tweet times of a user – drawn from a Poisson process of *time-varying rate*

- Rate during non-game times (λ_0 tweets/minute) (personal baseline)
- Rate during game times (λ_1 tweets/minute)

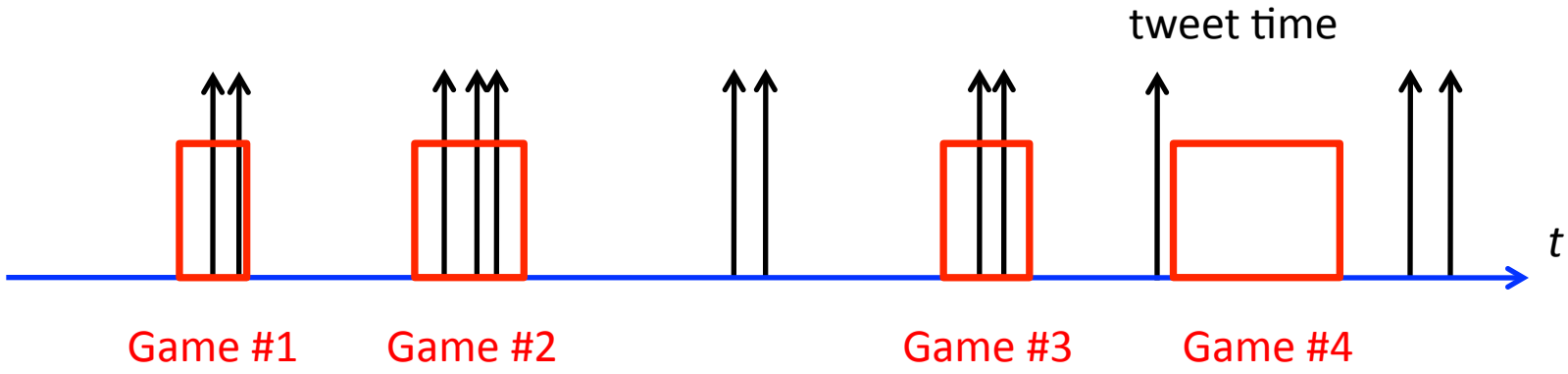
Model: Tweet times of a *Fan*



A fan tweets more often during game times $\lambda_1 > \lambda_0$

Statistic for fandom

- *Evidence*: Tweet times of user



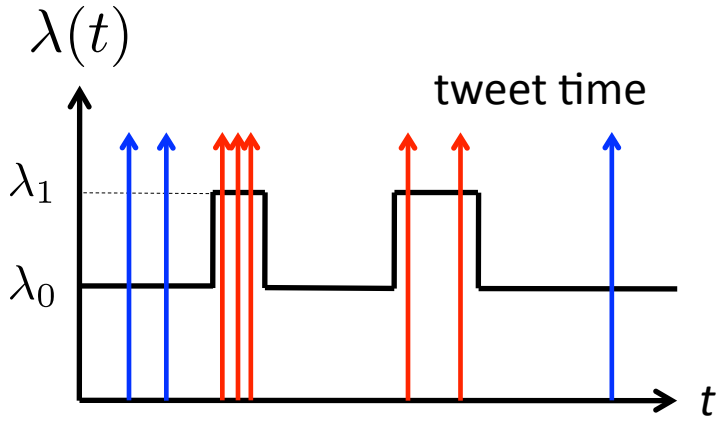
- Statistic for fandom:
  - How confident are we in the assertion that he/she has tweeted more often during games?

$$Z = \Pr\left[\lambda_1 > \lambda_0 | \text{Tweet times}\right]$$

Captures intuition missing in ratio of rates

Low sensing overhead

- Minimalistic model
  - Poisson with two rates:



- Sufficient statistic for λ_1 , λ_0
  - #Tweets during games N_1
  - #Tweets in non-game times N_0

- Exact tweet times not needed
- Easy to compute

$$Z = \Pr[\lambda_1 > \lambda_0 | \text{Tweet Times}]$$
$$= \Pr[\lambda_1 > \lambda_0 | N_0, N_1]$$

Low false alarm rates

False alarms: Proportion of non-fans misclassified as fans

- Particular interest
  - Fraction of "fans" is small (2%)
  - Most users are "non-fans" (98%)

- Moderate false alarm rates are bad! (5%)
  - Pool of users who clear the threshold
    Miss-classified non-fans (4.9%) >> Total fans (2%)

Detection rate 100% does not help

- Need low false alarms

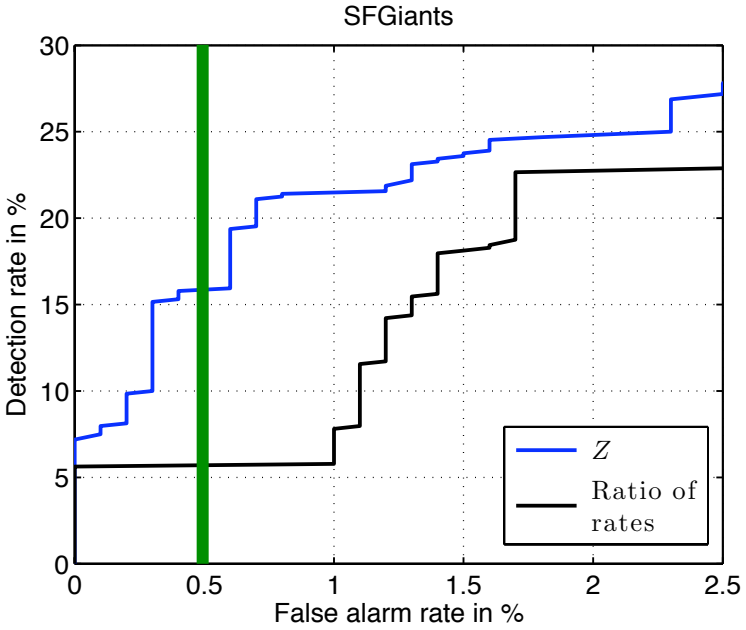
Dataset

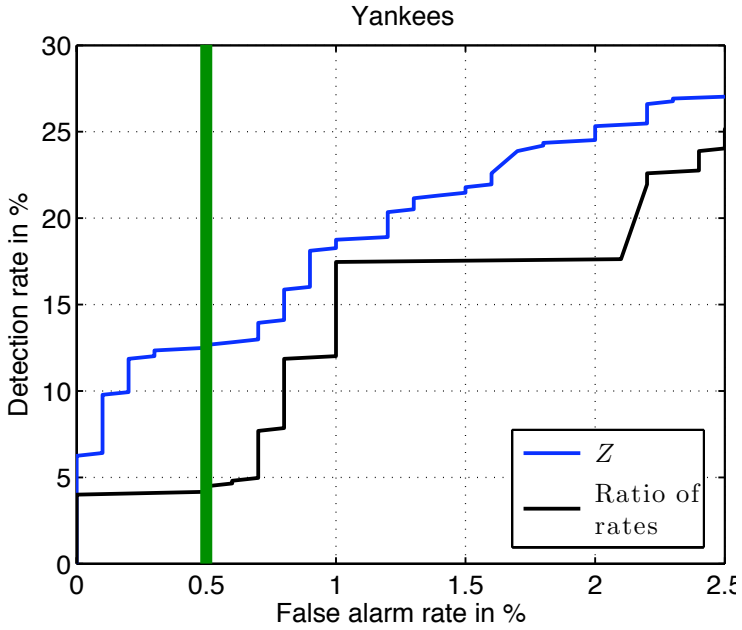
Predict fandom of SFGiants, Yankees

- Dataset
  - 10% random under-sampling – one month window
- Identified $\sim$ 600 fans
  - text analysis of
    { tweets in 15 minute window before & after each game }
- 1000 non-fans
  - randomly picked users

Results

Obtained by progressively *decreasing* the thresholds





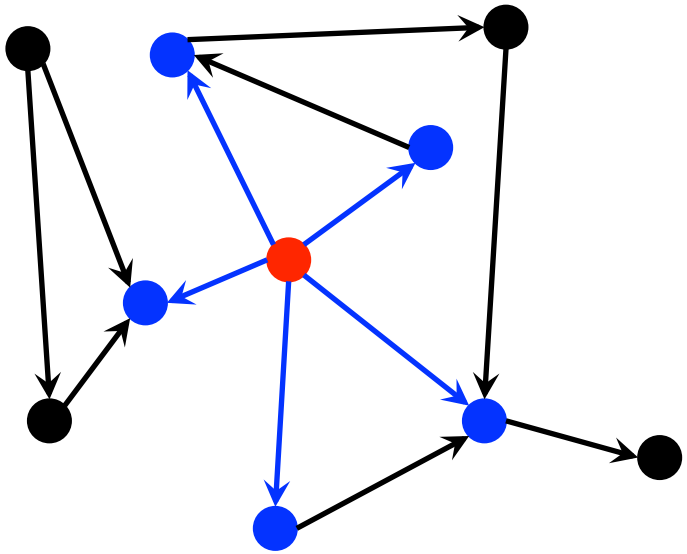
0.5% false alarm	RATIO OF RATES	BAYESIAN Z
SFGiants	5.6%	15.8%
Yankees	4.5%	12.7%

Minimal sensing and computation

No computations : Toss a coin → **false alarm = detection rate**

INFORMATION FROM NEIGHBORS

User interactions

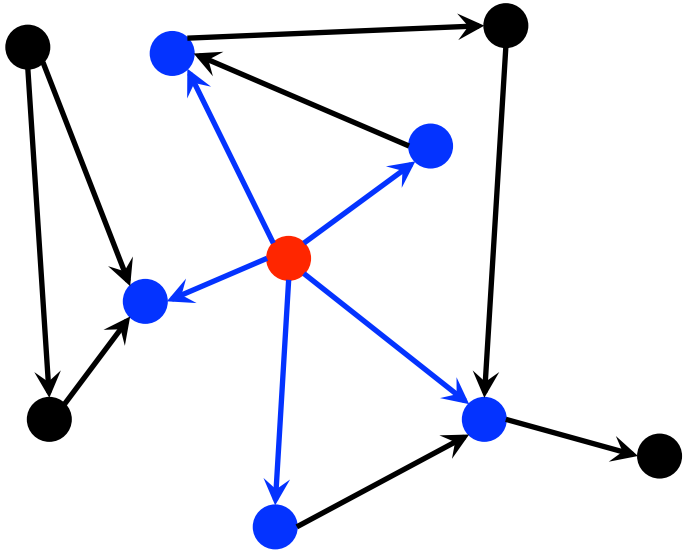


People you interact with are *more likely* to be fans if you are a fan

IMPROVE INFERENCE

Use tweet time dynamics of {neighbor} to *refine* estimates of your fandom

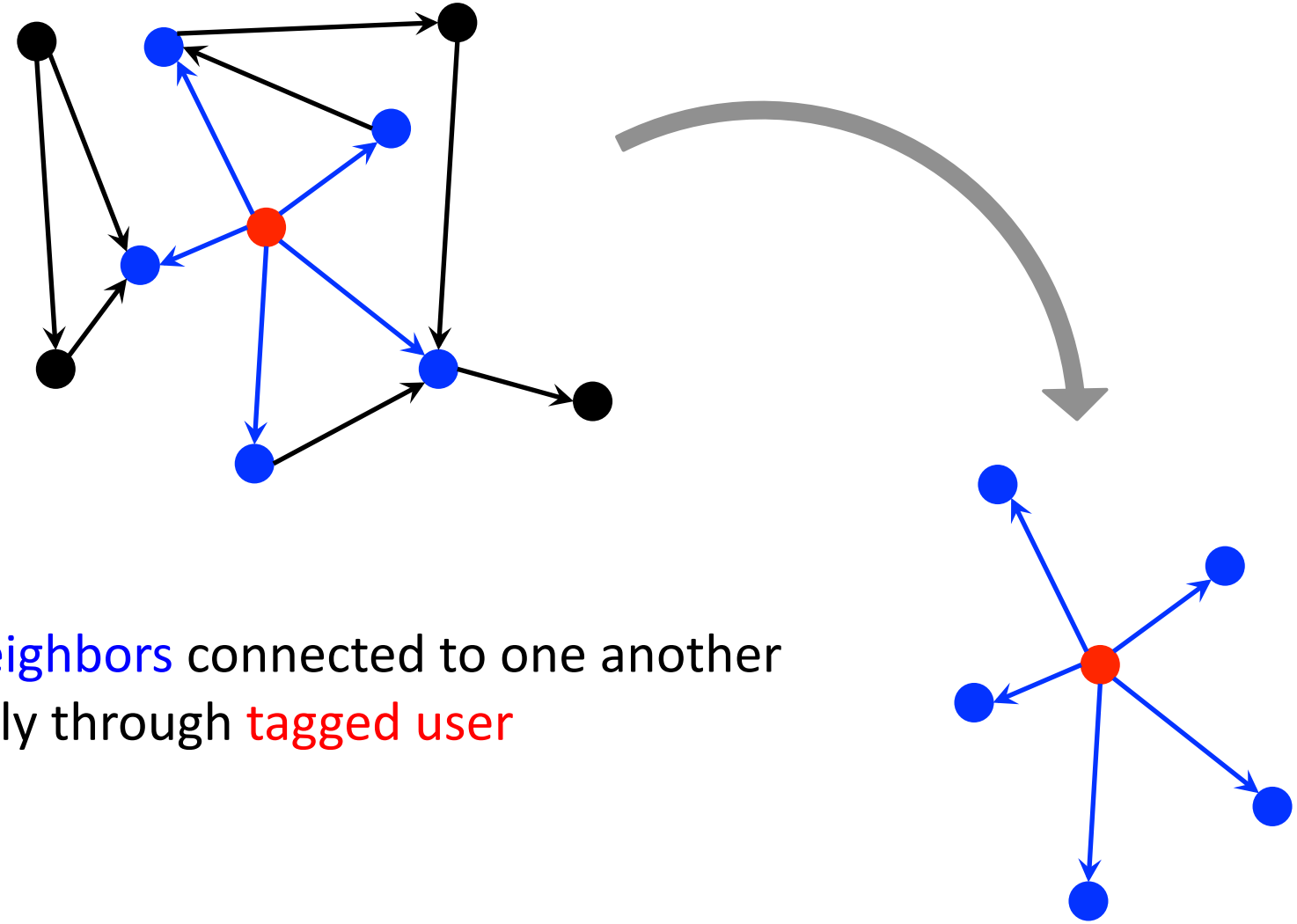
User interactions



Neighbors – {user handles referred to by tagged user during observation window}

- Available in *tweet meta-data*
  - Build from a *stream of tweets*
  - No need to parse tweet
- Captures *live* interactions

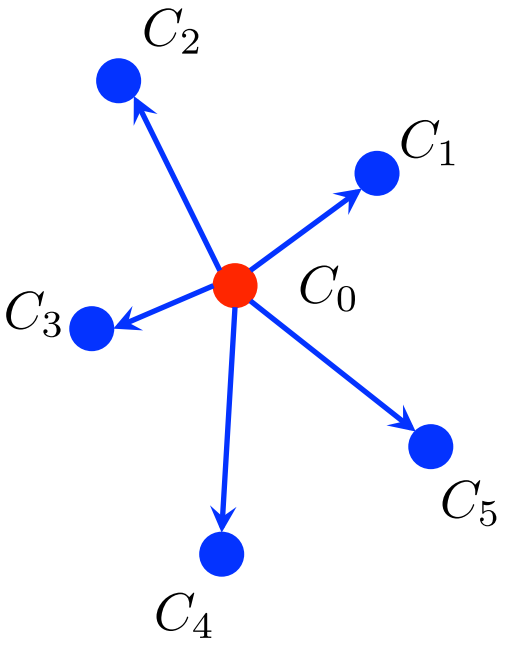
Simplified neighborhood



Neighbors connected to one another only through tagged user

Markov model

- Conditioned on the tagged user's fandom C_0, neighbor fandom $\{C_k\}$s are *independent* random variables



$$\Pr[C_1, \ldots, C_5 | C_0] = \prod_k \Pr[C_k | C_0]$$

Neighbor of a fan – more likely to be a fan than neighbor of not-a-fan

$$\alpha = \Pr[C_k = 1 | C_0 = 1]$$
$$\beta = \Pr[C_k = 1 | C_0 = 0]$$
$$\alpha \gg \beta$$

Not all users are timely

- A fan may not tweet during games
- Non-fan may tweet heavily during games (other interests?)

$p_t = \Pr[Y_k = 1 | C_k = 1]$ Probability of *fan being timely*

$p_f = \Pr[Y_k = 1 | C_k = 0]$ Probability of *false alarms* (non-fan tweeting aggressively during games)

Saturate per-user likelihoods

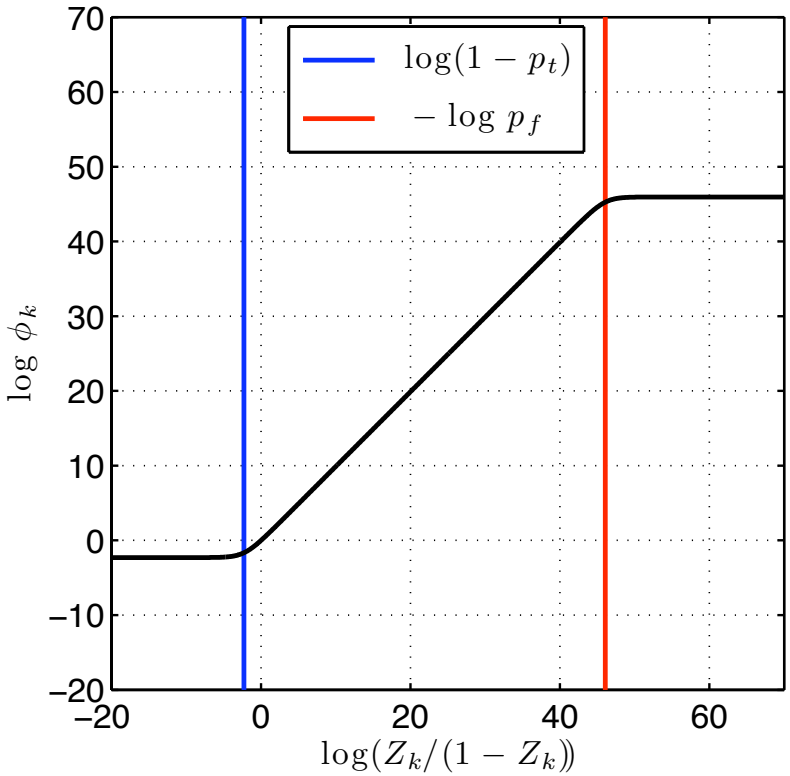
$$Z_k = \Pr[\lambda_1(k) > \lambda_0(k) | \text{Tweet times}]$$
$$= \Pr[Y_k = 1 | \text{Tweet times}]$$

Soft-thresholds per-user likelihood ratios:
$Z_k \,/\, (\,1 - Z_k\,)$

Acknowledge that Z_k makes mistakes

$$\phi_k = \frac{\Pr[\text{Tweet times}|C_k = 1]}{\Pr[\text{Tweet times}|C_k = 0]}$$
$$= \frac{1 + p_t\left(\frac{Z_k}{1-Z_k} - 1\right)}{1 + p_f\left(\frac{Z_k}{1-Z_k} - 1\right)}$$

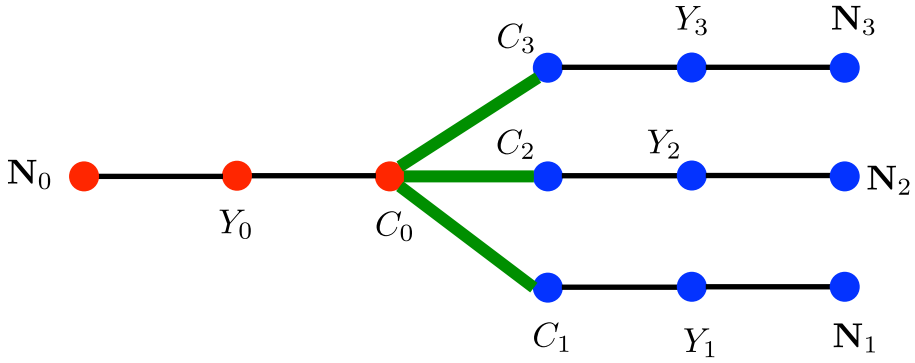
$p_t = 0.9$
$p_f = 10^{-20}$



Consolidated statistic

Fuse all observations

Statistic is Log Likelihood Ratio of all observations



$$S = \log \frac{\Pr[\mathbf{N}_0, \mathbf{N}_1 \ldots, \mathbf{N}_L | C_0 = 1]}{\Pr[\mathbf{N}_0, \mathbf{N}_1 \ldots, \mathbf{N}_L | C_0 = 0]}$$

$$= \log \phi_0 + \sum_{n=1}^{n=L} \log \underbrace{\frac{1 + \alpha\,(\phi_n - 1)}{1 + \beta\,(\phi_n - 1)}}$$

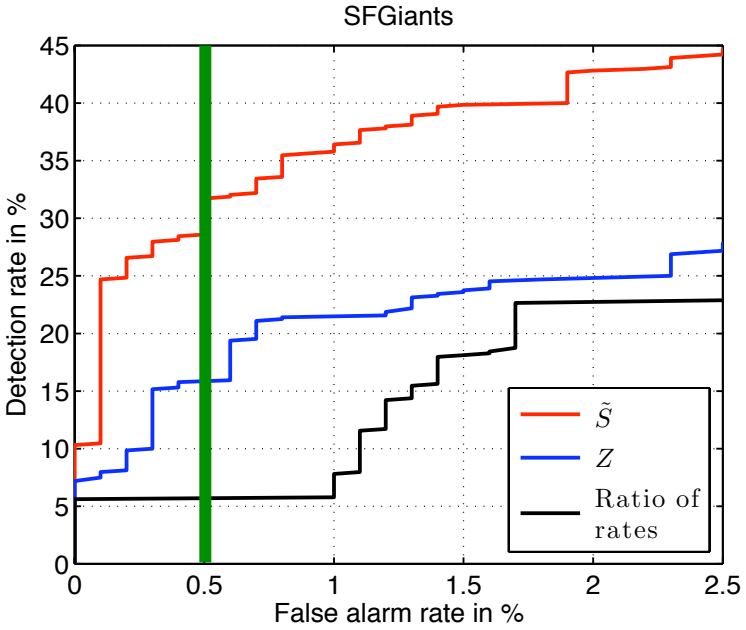
Further saturate neighbor likelihood ratios

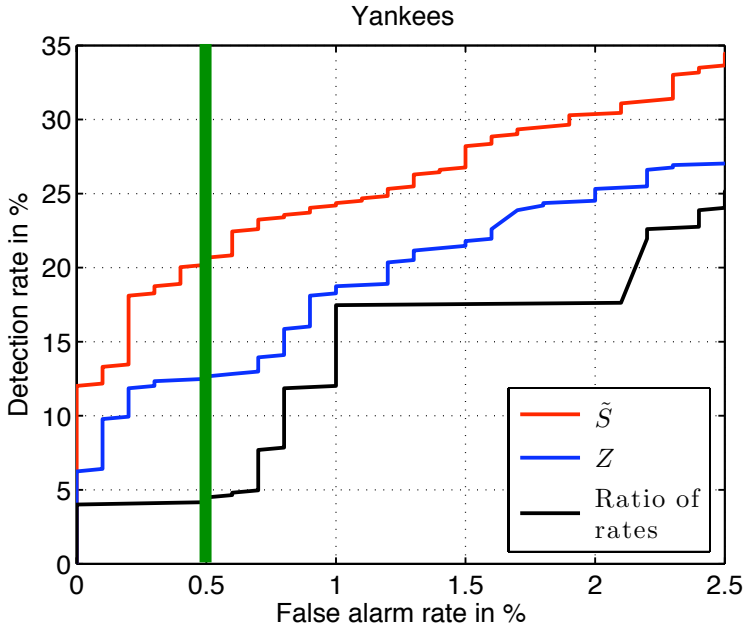
Simplify (reduce parameters)

$$\tilde{S} = \log \phi_0 + \kappa \sum_{n=1}^{n=L} \log \phi_n$$

Results

$$\tilde{S} = \log \phi_0 + \kappa \sum_{n=1}^{n=L} \log \phi_n$$ Parameters ($p_f = 10^{-20}$, $p_t = 0.9$, $\kappa = 1/6$)

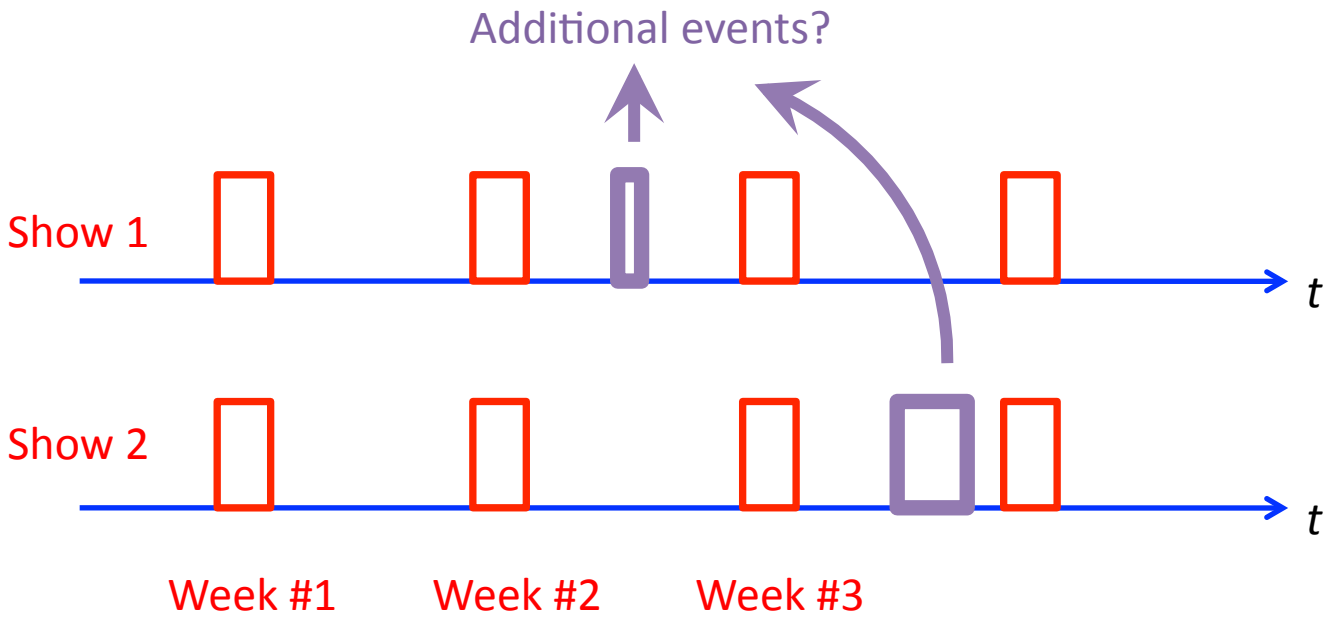




0.5% false alarm	RATIO OF RATES	BAYESIAN Z	USER + NEIGHBORS $\tilde{S}$
SFGiants	5.6%	15.8%	31.7%
Yankees	4.5%	12.7%	20.7%

Limitations

- Heavy event times overlap among different interests

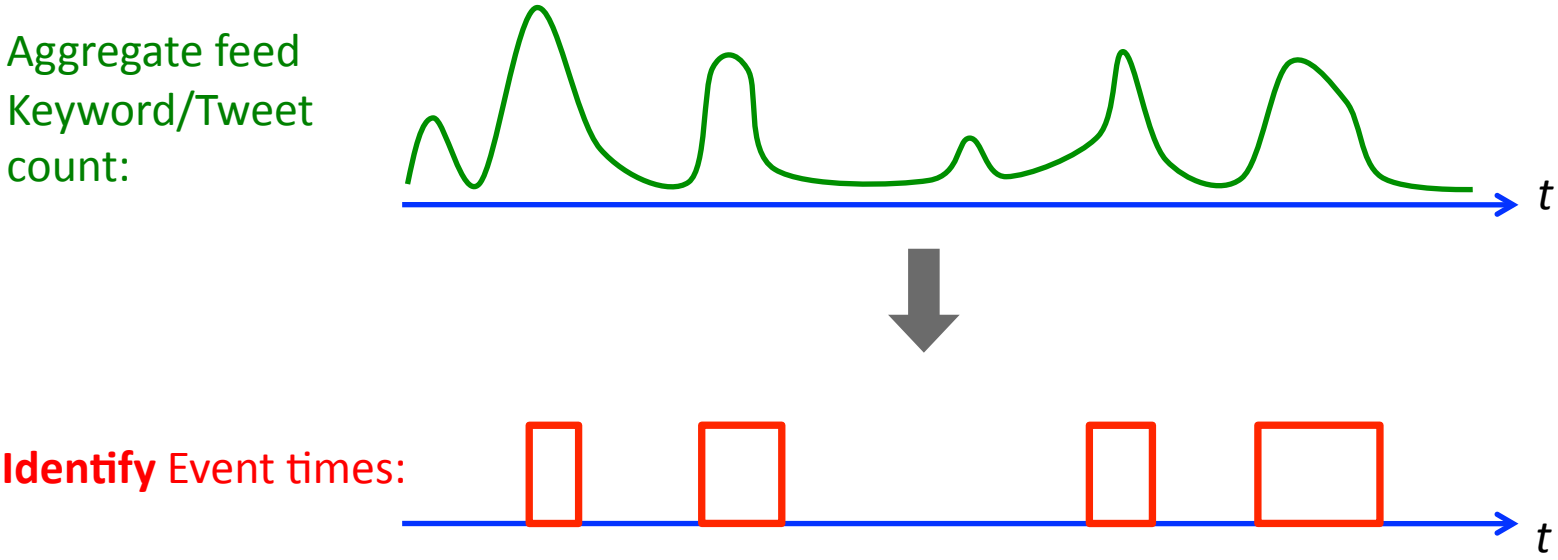


"**How different**" should two event windows be?

- Interests must elicit a timely response from users

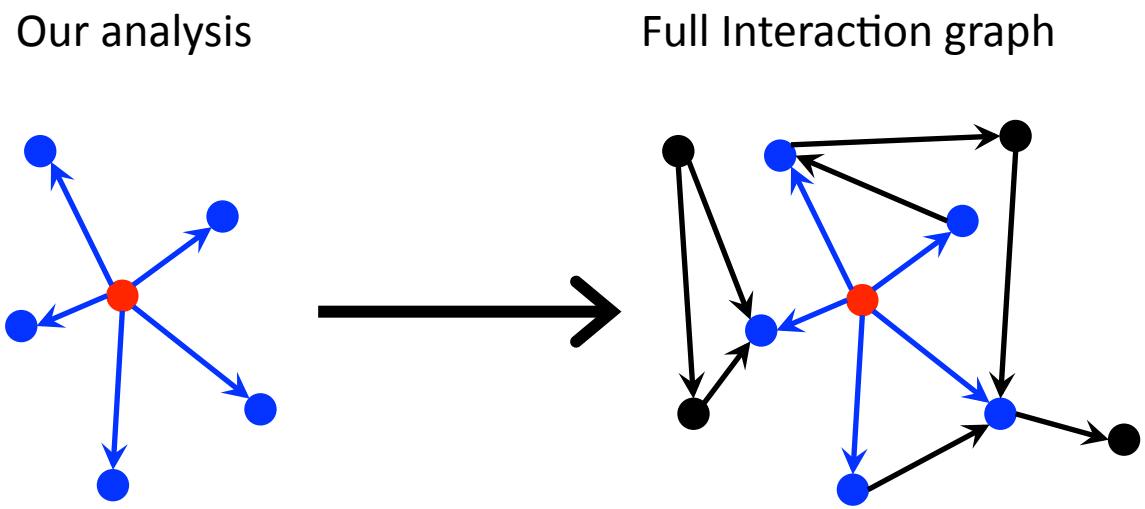
Future work

- *Learn topic-specific event times* in a data-driven manner?
  - Run text analysis on aggregate feeds? News feeds, etc
    - **Associate topics with time**
  - Feeds can be *targeted to the topic*



Future work

- Neighbor interactions overlap
  - Extend the Bayesian approach



- Strength of interactions?

Conclusion

- Value to using times of tweets/posts
  - Good detection performance at low false alarm rates
  - Scalable: low sensing, computational overhead
  - Complement existing methods
- Interactions provide a lot of information
  - Further improves detection accuracy
- Interesting directions for future research
  - Experiments to identify interests – timely response
  - Learn event times data-driven manner
  - Incorporate graph structure

Thank you