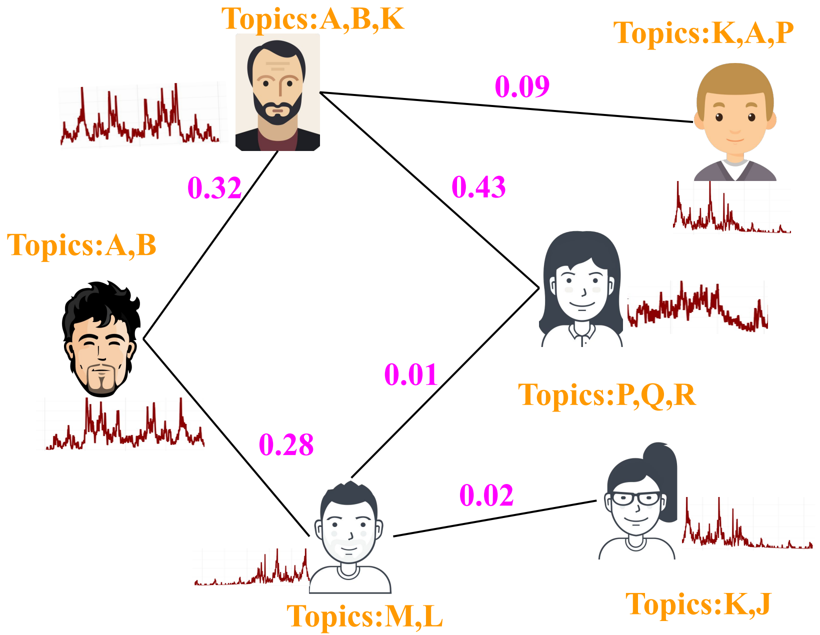
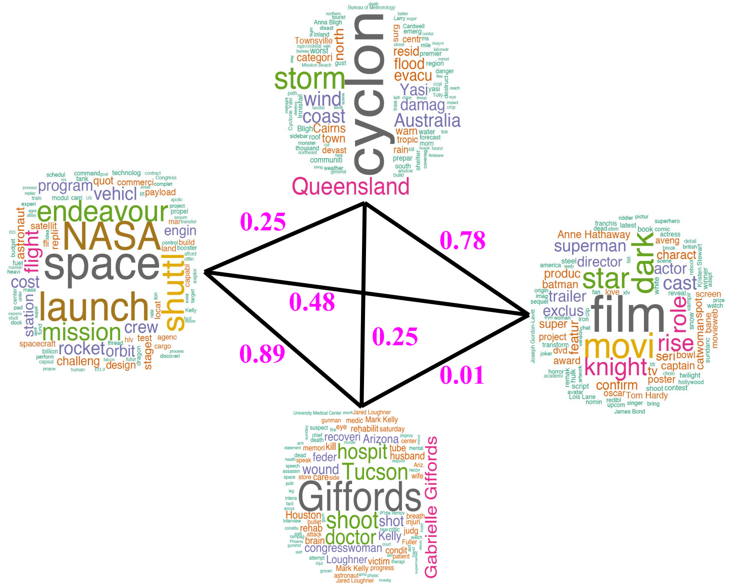
Discovering Topical Interactions in Text-based Cascades using Hidden Markov Hawkes Process

Srikanta Bedathur (IIT Delhi), Indrajit Bhattacharya (TCS Research),
Jayesh Choudhari, Anirban Dasgupta (IIT Gandhinagar)

Motivation

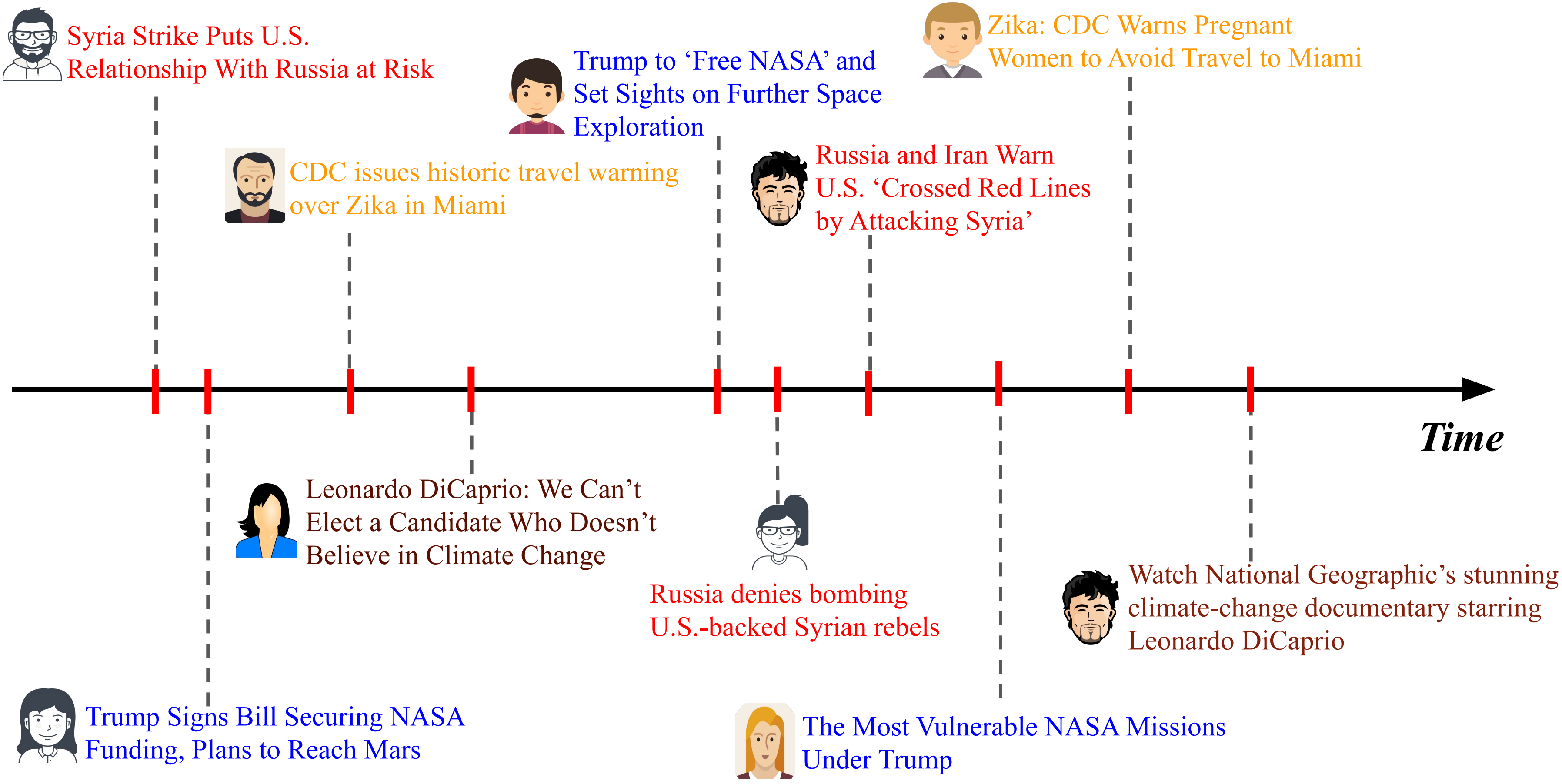




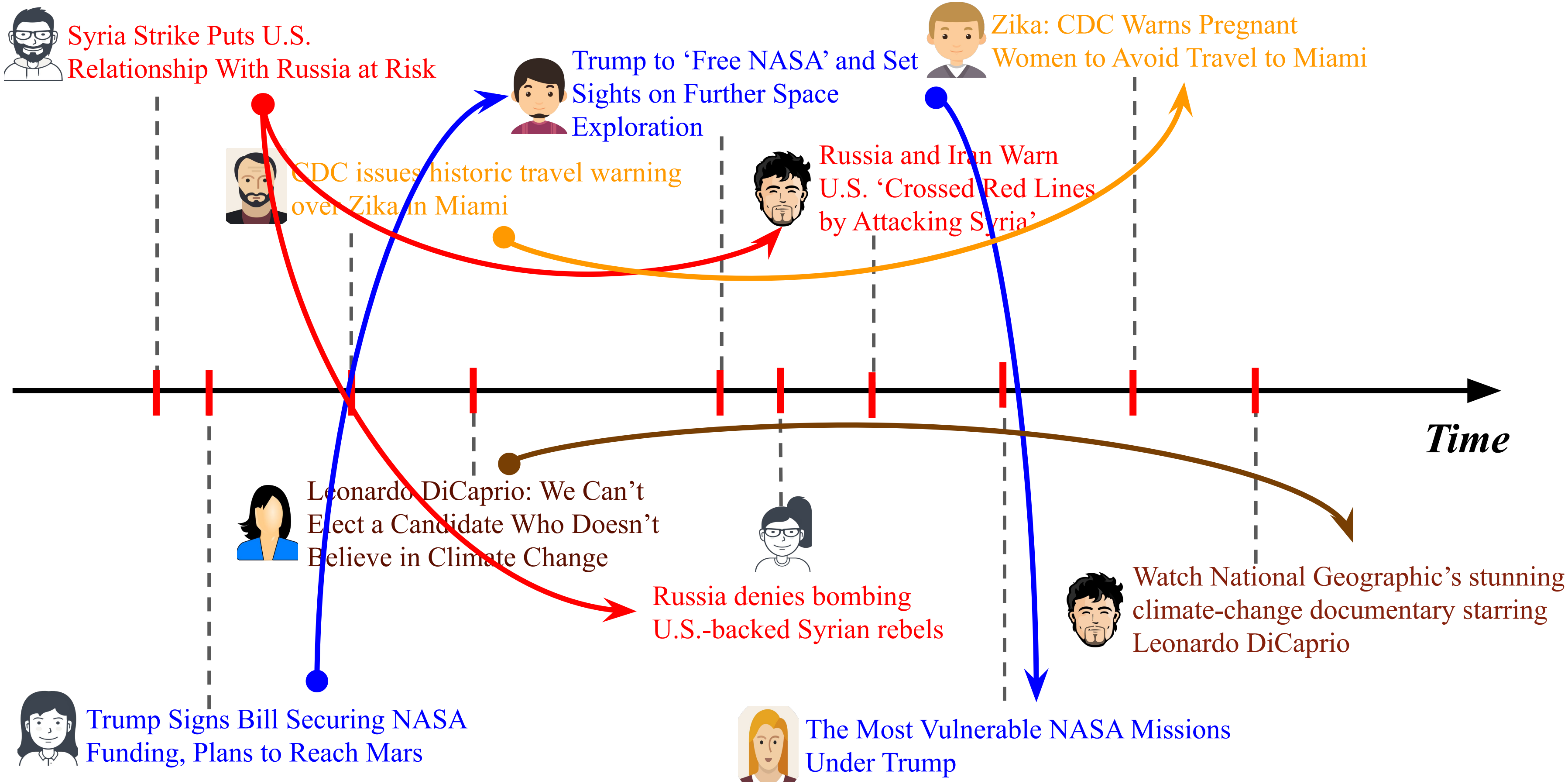
- User Temporal Dynamics
- Preferred topics of each user
- Network Strengths (user-user influence)

- Topics
- Topical Interactions

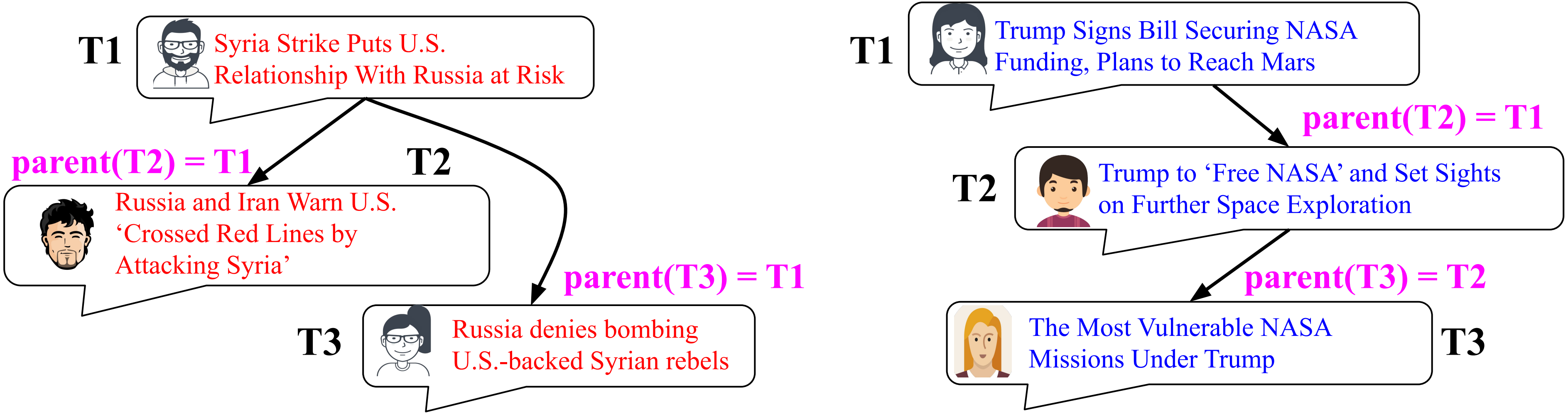
Data: Network + Time-series of Tweets



Mixture of Conversations



Cascades (Separate Conversations)



Just separate this conversations out!!!

Hidden Markov Hawkes Process

- Coupling of Network (Multivariate) Hawkes Process and the Markov Chain over topics.

- Coupled inference: Collapsed Gibbs sampling

Snapshot of Results

Parent-Child tweet pair

Gellman:My definition of whistleblowing:are you shedding light on crucial decision that society should be making for itself. #snowden

Gellman we are living inside a one way mirror,they & big corporations know more and more about us and we know less about them #sxsw

Why Topical Interactions?

Hashtags from top-3 transitioned topics

agentsofshield, arrow, tvtag, supernatural, chicagoland

Topic-1: *idol, bbcan2, havesandhavenots, thegamebet*
Topic-2: *tvtag, houseofcards, agentsofshield, arrow,*
Topic-3: *soundcloud, hiphop, mastermind, nowplaying*

Hashtags from a pair of parent-child topics

steelers,browns,seahawks, fantasyfootball, nfl

mlb, orioles, rays, usmnt, redsox

Generative Model

HMHP Generative Process

1) Generate (t_e, c_e, z_e) for all events according Multivariate Hawkes Process.
2) For each topic k: sample $\boldsymbol{\zeta}_k \sim Dir_{\mathcal{W}}(\boldsymbol{\alpha})$
3) For each topic k: sample $\boldsymbol{\mathcal{T}}_k \sim Dir_K(\boldsymbol{\beta})$
4) For each node v: sample $\boldsymbol{\phi}_v \sim Dir_K(\boldsymbol{\gamma})$
5) For each event e at node $c_e = v$:
   a) i) **if** $z_e = 0$ (level 0 event):
      draw a topic $\eta_e \sim Discrete_K(\boldsymbol{\phi}_v)$
      ii) **else**:
      draw a topic $\eta_e \sim Discrete_K(\boldsymbol{\mathcal{T}}_{\eta_{z_e}})$
   b) Sample document length $N_e \sim Poisson(\lambda)$
   c) For $w = 1 \ldots N_e$: draw word $x_{e,w} \sim Discrete_{\mathcal{W}}(\boldsymbol{\zeta}_{\eta_e})$

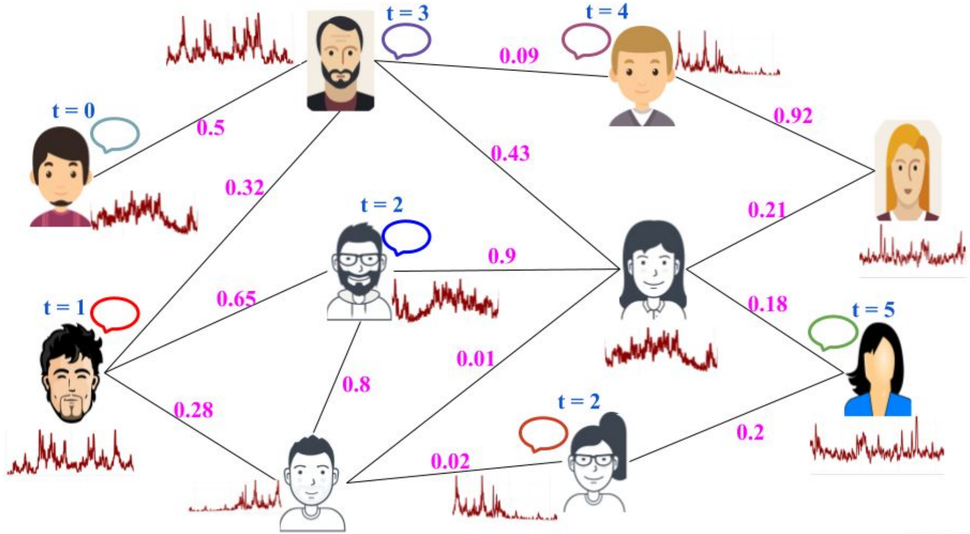
Temporal Dynamics and Network Inference using Multivariate Hawkes Process

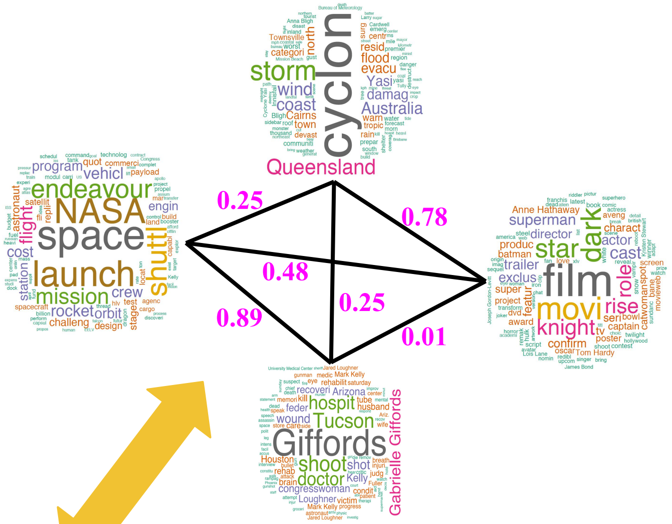
Cascade reconstruction and Topical Interactions coupling Multivariate Hawkes Process and Topical Markov Chains

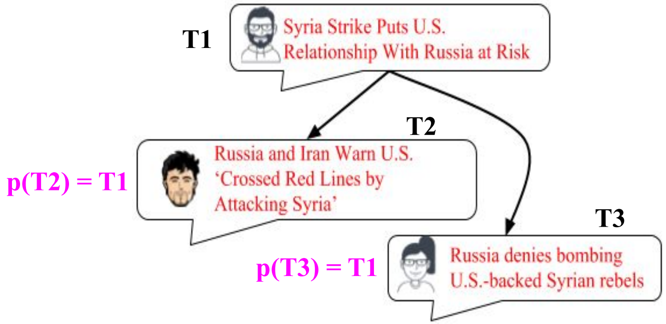
Topic Model

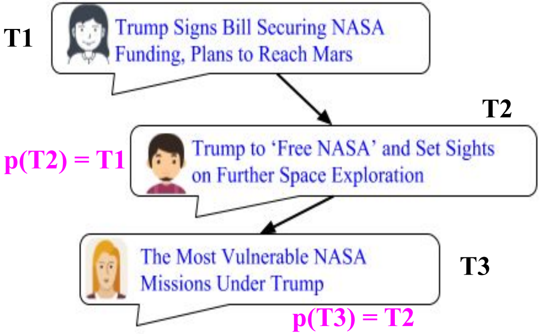
Inference

Challenge - Coupled Problems



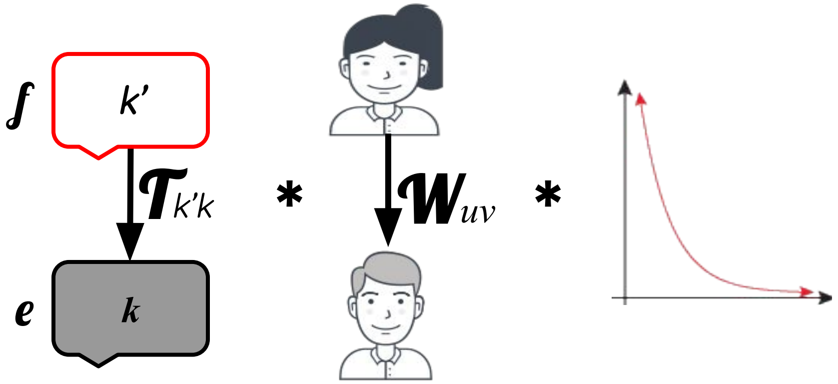






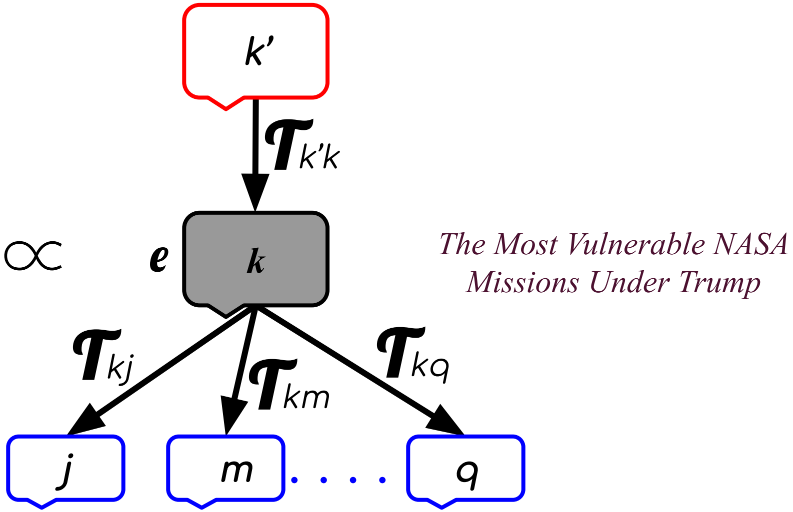
Cascade Inference

$$\mathcal{P}\left[par(e) = f \mid Topics, W, \mu, timeStamps \right] \propto$$



Topic Inference

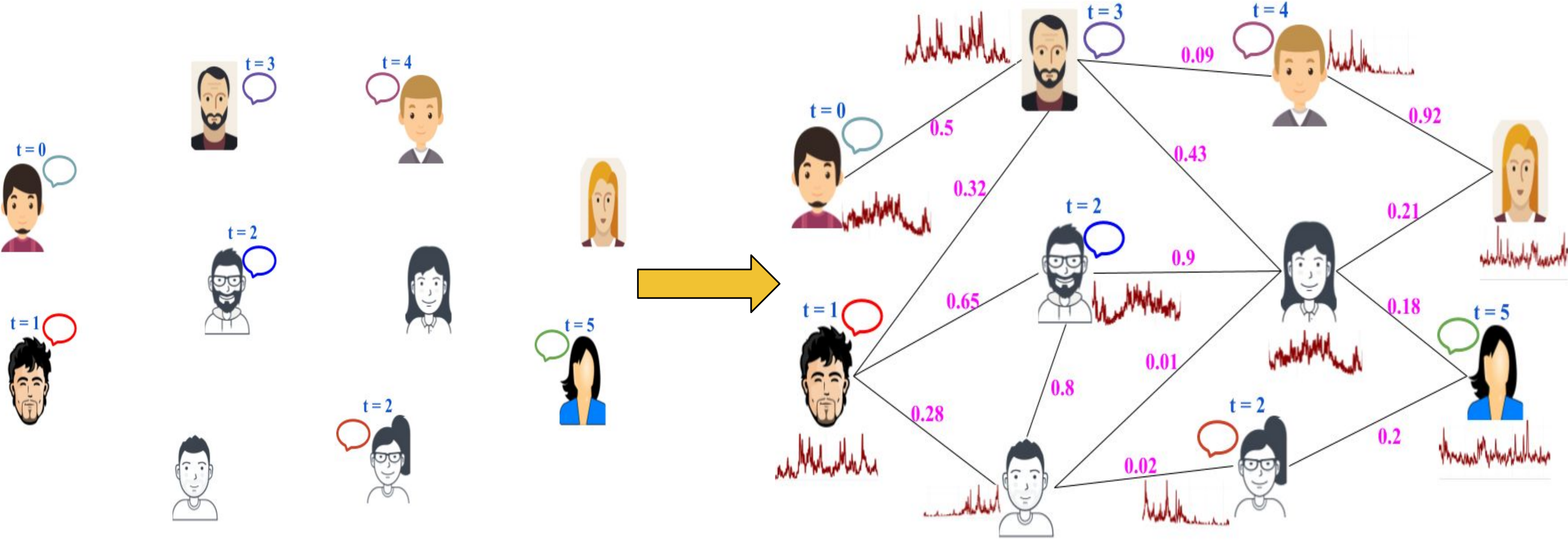
$$\mathcal{P}\left[\, topic(e) = k \;\middle|\; parentStructure,\ tweetText,\ \{topic(f) \mid f \mathrel{!=} e\} \,\right] \propto$$



Note: Topical Interactions are inferred using the sampled topics and the parent-child structure

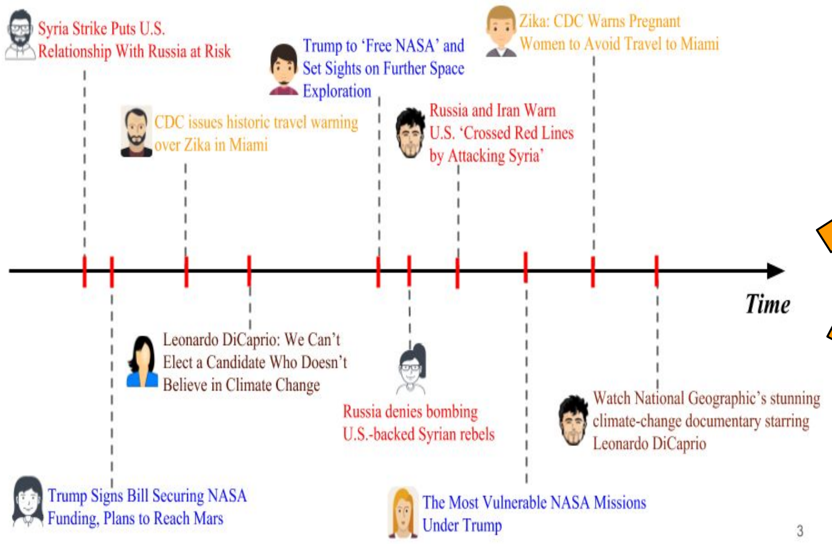
Existing Models

Network Hawkes Model

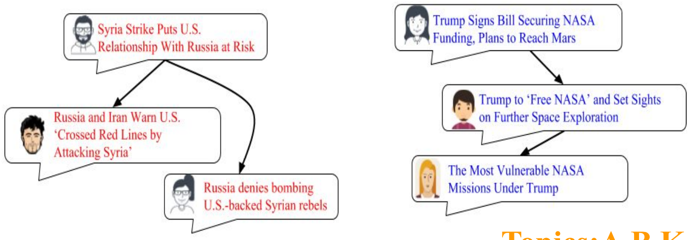


Does not model (textual) content of events / tweets

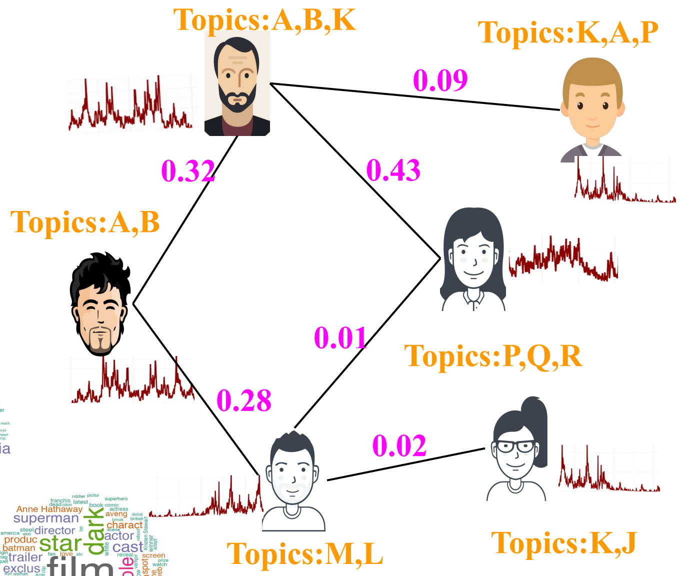
Hawkes Topic Model [He et al. ‘15]



Set of tweets and time-stamps



Cascade Reconstruction



Topic Model

Network Reconstruction and Temporal Dynamics

Missing Topical Interactions in HTM

[#MASalert] Statement By Our Group CEO, Ahmad Jauhari Yahya on MH370 Incident. Released at 9.05am/8 Mar 2014

Missing #MalaysiaAirlines flight carrying 227 passengers (including 2 infants) of 13 nationalities and 12 crew members.

Repeating patterns in the topics of the parent and child events

Generation of Topic of child event in HTM

*If event **e** is not spontaneous, then*

$$\mathbf{Topic}(e) \sim \mathbf{Normal}(\mathbf{Topic}(\mathbf{parent}(e)), \sigma^2 I)$$

v/s

Generation of Topic of child event in HMHP

*If event **e** is not spontaneous, then*

$$\mathbf{Topic}(e) \sim \boldsymbol{\zeta}(\mathbf{Topic}(\mathbf{parent}(e)))$$

where, $\boldsymbol{\zeta}$ is Topical Interaction Distribution

Note: These parent-child pairs are neither retweets nor does twitter provide any signal to know any relation about these pairs

Results

Datasets

Twitter (Real Data):

- ***500K tweets*** *corresponding to top 5K hashtags from the most prolific 1M users generated in a contiguous part of March 2014*

Semi-Synthetic:

- *Retain the underlying set of nodes and the follower graph from a sample of Twitter Data.*

- *Estimate the parameters required for our model from the data.*

- *Generate* ***5 different samples*** *of* ***1M events*** *using* ***HMHP*** *model.*

Baselines

- ***HWK + DIAG:***
  - *Simplified HMHP with diagonal topical interactions*
- ***HWK x LDA:***
  - *Network Hawkes model for cascade structure and time-stamps*
  - *LDA mixture model for the textual content*
- ***HTM (Hawkes Topic Model)***

Reconstruction Accuracy *(Semi-Synthetic Dataset)*

	HMHP	HWK+Diag	HWK×LDA
Mean APE	**0.448**	0.565	0.552
Median APE	**0.255**	0.283	0.287

Network Reconstruction Error

Mean Error :- ~18% lower
Median Error :- ~10% lower

	HMHP	HWK+Diag	HWK×LDA
Accuracy	**0.581**	0.362	0.37
Recall@1	**0.595**	0.373	0.38
Recall@3	**0.778**	0.584	0.589

Cascade Reconstruction Accuracy

Acc/Recall@1 :- ~57% better
Recall@3 :- ~32% better

Topic	HMHP	HWK+Diag	HWK×LDA
Precision	**0.893**	0.123	0.781
Recall	**0.746**	0.367	0.752
F1	**0.811**	0.18	0.765

Topic Identification

HMHP performs ~5-6% better

Generalization Performance *(Twitter Dataset)*

Heldout Log-Likelihood

#Topics	Log-Likelihood	HMHP	HWK + Diag	HWK x LDA
25	*Content*	*-30499278*	*-33356945*	*-30532938*
	Time	*-4236958*	*-4042903*	*-4299630*
	Total	***-34736237***	*-37399849*	*-34832568*
50	*Content*	*-30141081*	*-33427354*	*-30089733*
	Time	*-4288438*	*-4510072*	*-4343571*
	Total	***-34429519***	*-37937426*	*-34433305*
75	*Content*	*-29860909*	*-33433922*	*-29861050*
	Time	*-4285293*	*-4510535*	*-4373736*
	Total	***-34146202***	*-37944457*	*-34234787*

HMHP performs ~5% better than the baselines

Summary

- *Generative model for textual time-series from user networks having topical interactions*

- *Couples Topical Markov Chains and Multivariate Hawkes Processes*

- *Scalable collectively inference using collapsed Gibbs Sampling*

- *More accurate cascade reconstruction, topic identification and network reconstruction and better generalization for test data*

- *Derive insights about topical interactions that the existing models cannot*

Thank You