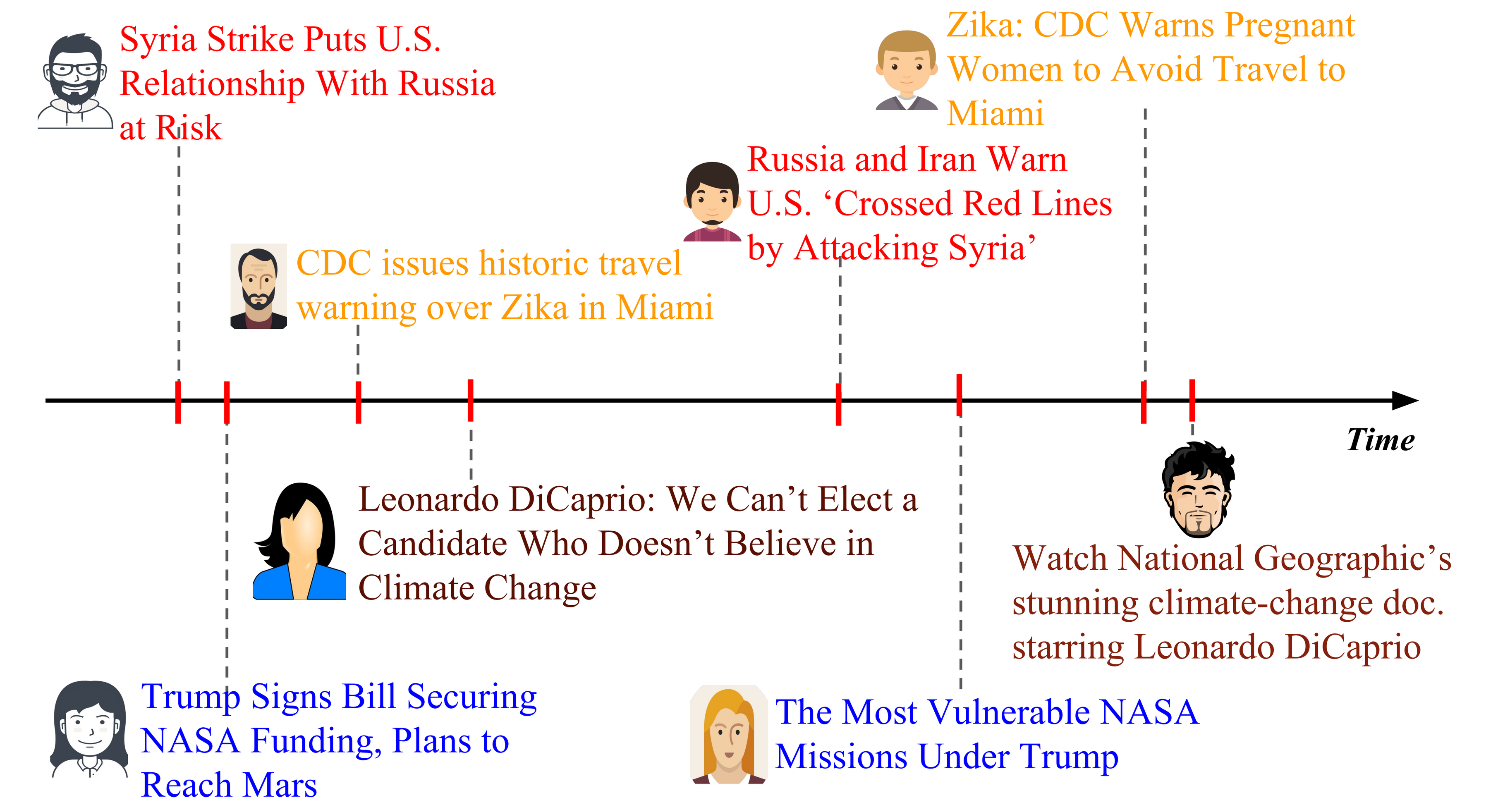
Discovering Topical Interactions in Text-based Cascades using Hidden Markov Hawkes Process (HMHP)

Srikanta Bedathur[1], Indrajit Bhattacharya[2], Jayesh Choudhari[3], Anirban Dasgupta[3]

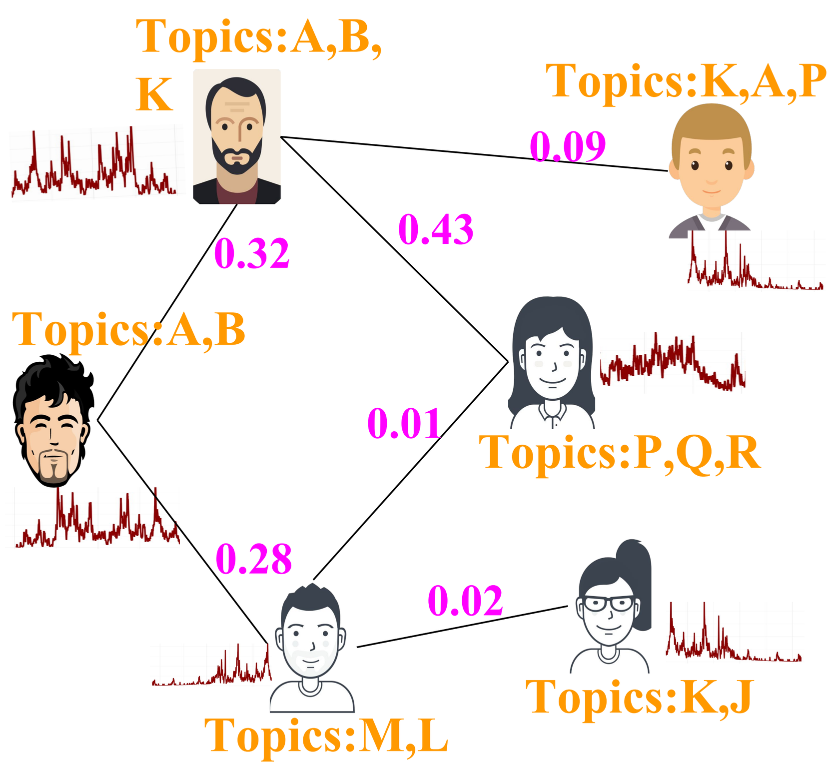
1. IIT Delhi, India 2. TCS Research Kolkata, India 3. IIT Gandhinagar, India

Motivation

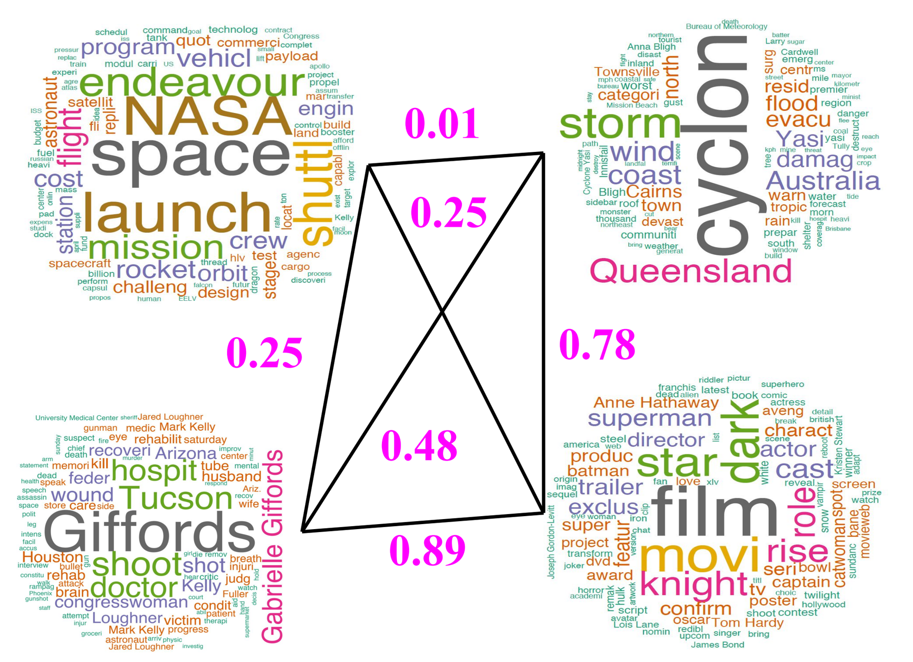


Time-series of Tweets (Mixture of conversations)

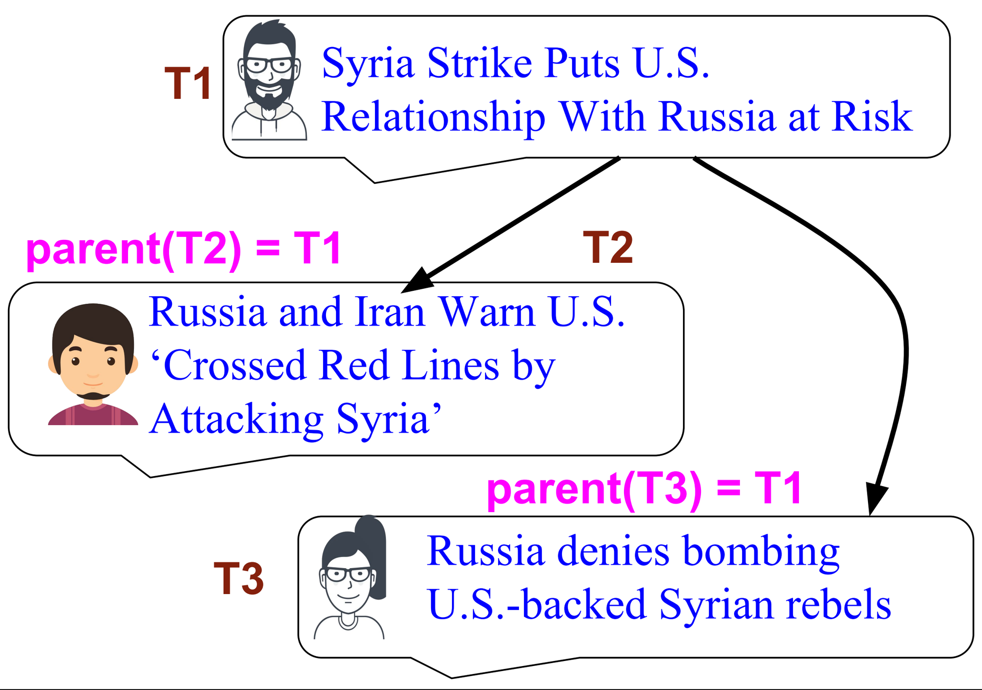
Questions



- *When and how frequently do users generate content and on what topic? **(Temporal Dynamics and Preferred Topics of each user)***
- *Who responds to whom and how quicky? **(User-User Influence).***
- *What are the different conversations i.e. Parent-Child Structure among Tweets? **(Cascade Reconstruction)***



- *What are the various Topics in the data and how do topics interact? **(Topics and Topical Interactions)***



Why Topical Interactions?

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

- ***Parent-child from different topics***
- ***Topic pair occurs frequently***
- ***HMHP assigns to different topics with high transition probability***

Hashtags from a pair of parent-child topics

steelers,browns,seahawks, fantasyfootball, nfl

mlb, orioles, rays, usmnt, redsox

Frequent topical transitions from football related hashtags to baseball related hashtags

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*

Random walk over topics to detect topic drifts - from tv shows to entertainment

HMHP Generative Model

- **Coupled Multivariate Hawkes Processes and (Hidden) Markov Chains**
- **Coupled inference: Collapsed Gibbs sampling**

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

Events are generated according to ***Multivariate Hawkes Process.***

Topic of event is sampled as one which is more related to or interacts with parents topic. **(Markov Chain over Topics)**

Repeating patterns in the topics of the parent and child events

[#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.

Generation of Topic of child event in HTM [1]

*If event **e** is not spontaneous, then*
$$Topic(e) \sim Normal(Topic(parent(e)), \sigma^2 I)$$

v/s

Generation of Topic of child event in HMHP

*If event **e** is not spontaneous, then*
$$Topic(e) \sim \mathcal{T}(Topic(parent(e)))$$
where, $\mathcal{T}$ is Topical Interaction Distribution

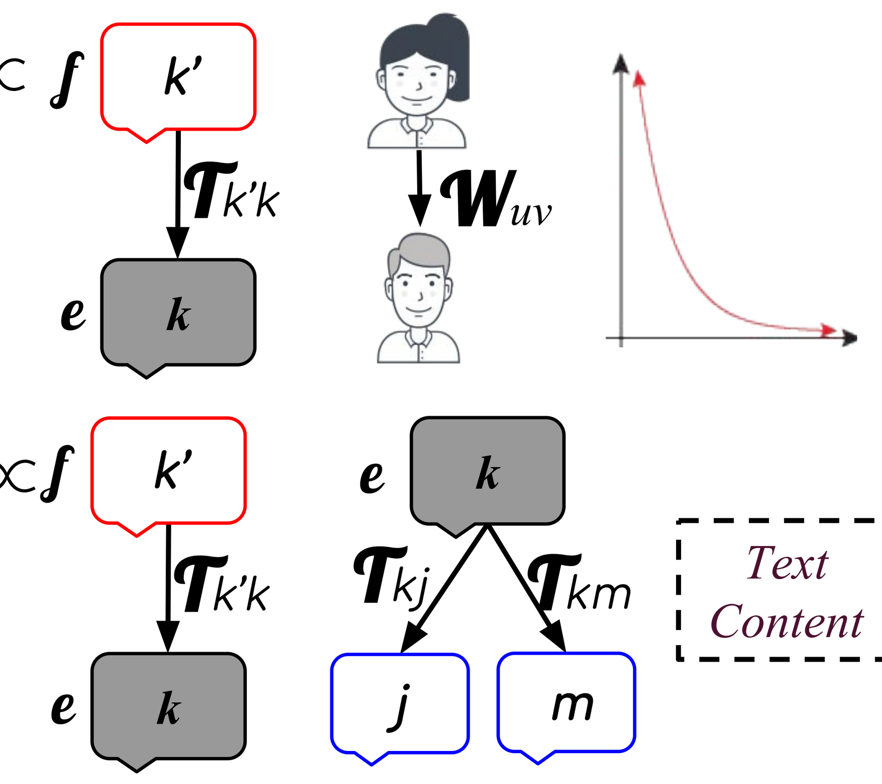
Inference

$$P(parent(e) = f \mid Topics, W, \mu, timeStamps) \propto f$$

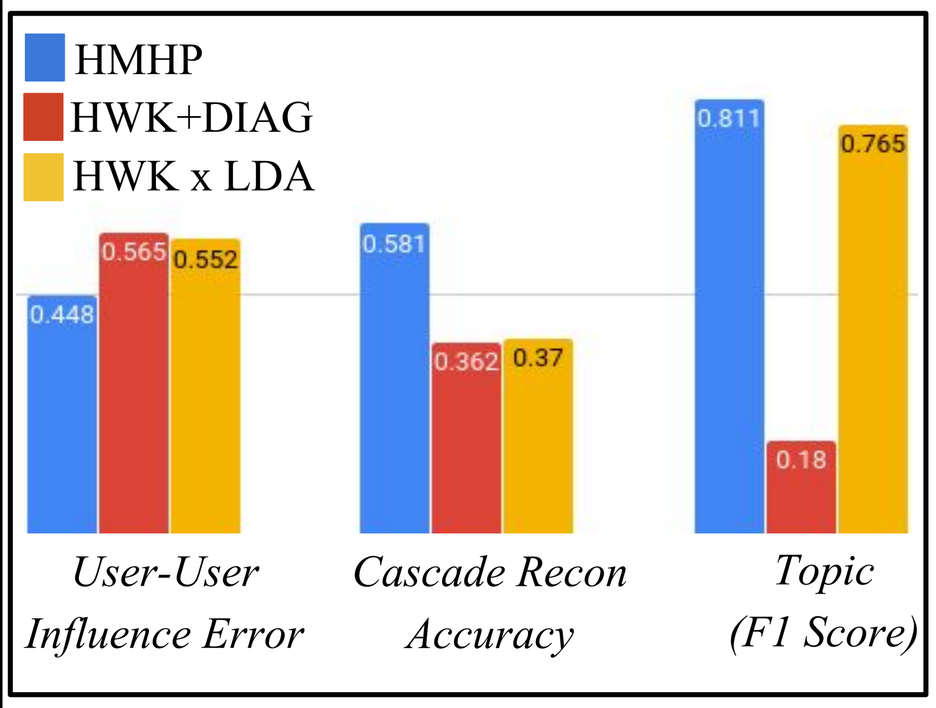
Probability of event f being a parent of event e is proportional to ***topical interaction*** between topic of event f and topic of event e.

$$P(Topic(e) = k \mid parents, tweet, \{Topic(f) \mid f \,!\!= e\}) \propto f$$

Probability of event e having topic k is proportional to ***topical interaction*** between the parents topic and topic k ***topical interaction*** between k topics of child events.



Results



Reconstruction Accuracy (Semi-Synthetic Data)

- **HWK + DIAG**: HMHP + diagonal Topic Interactions
- **HWK x LDA**: Networks Hawkes [2] + LDA Mixture Model (for content)

Heldout Log-Likelihood

#Topics	HMHP	HWK+Diag	HWKxLDA
25	**-34736237**	-37399849	-34832568
50	**-34429519**	-37937426	-34433305
75	**-34146202**	-37944457	-34234787

Generalization Performance (Twitter Data)

Significant improvement over HTM [1] on scaled down datasets. HTM [1] does not scale for our dataset.

References

1) He, X., Rekatsinas, T., Foulds, J., Getoor, L., & Liu, Y. (2015, June). Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In ICML
2) Linderman, S., & Adams, R. (2014, January). Discovering latent network structure in point process data. In International Conference on Machine Learning (pp. 1413-1421).