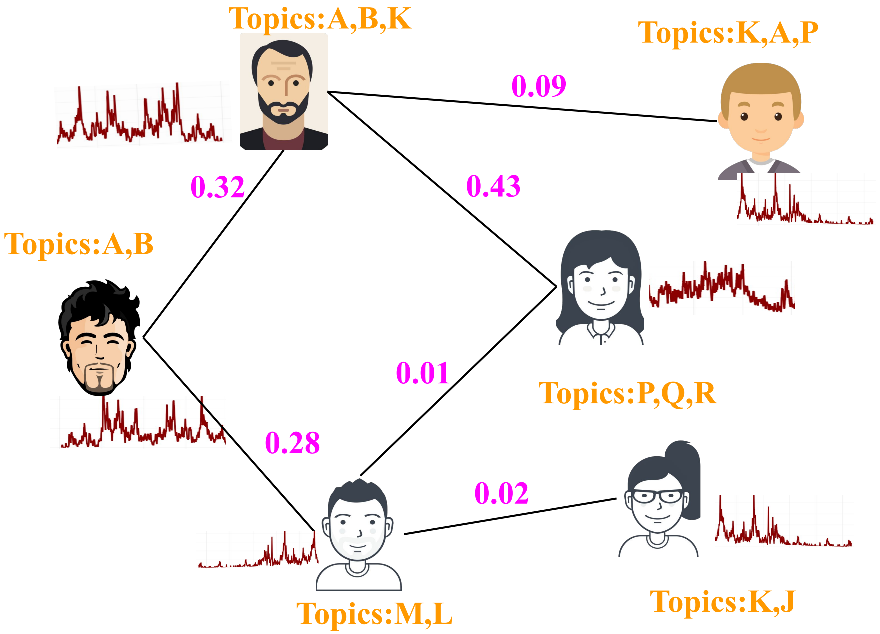
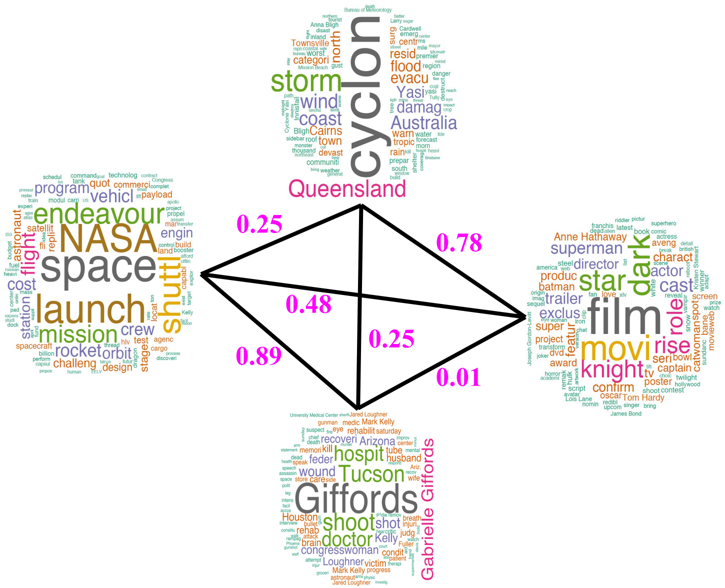
Discovering Topical Interactions in Text-based Cascades using Hidden Markov Hawkes Process

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

Motivation

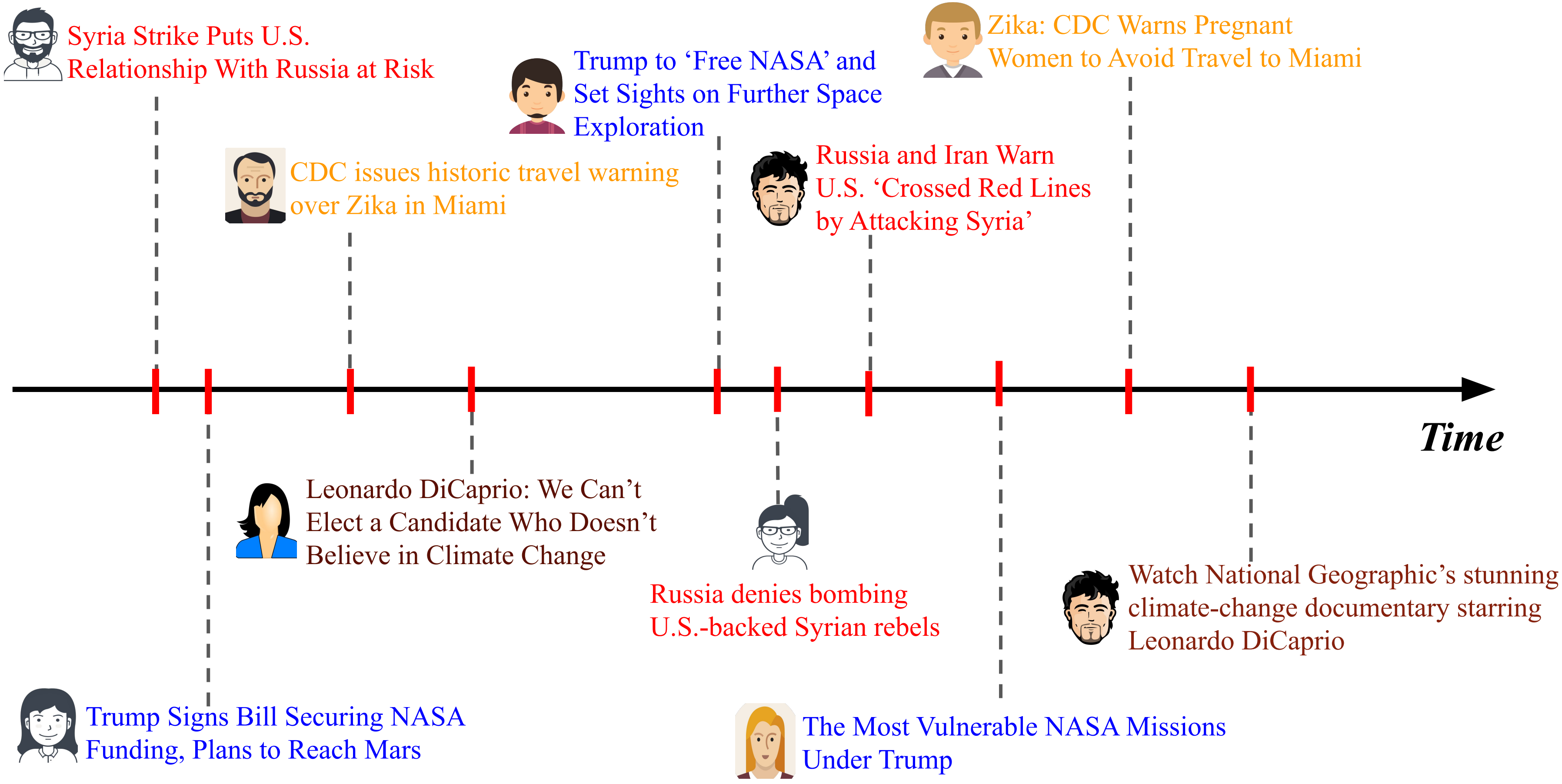




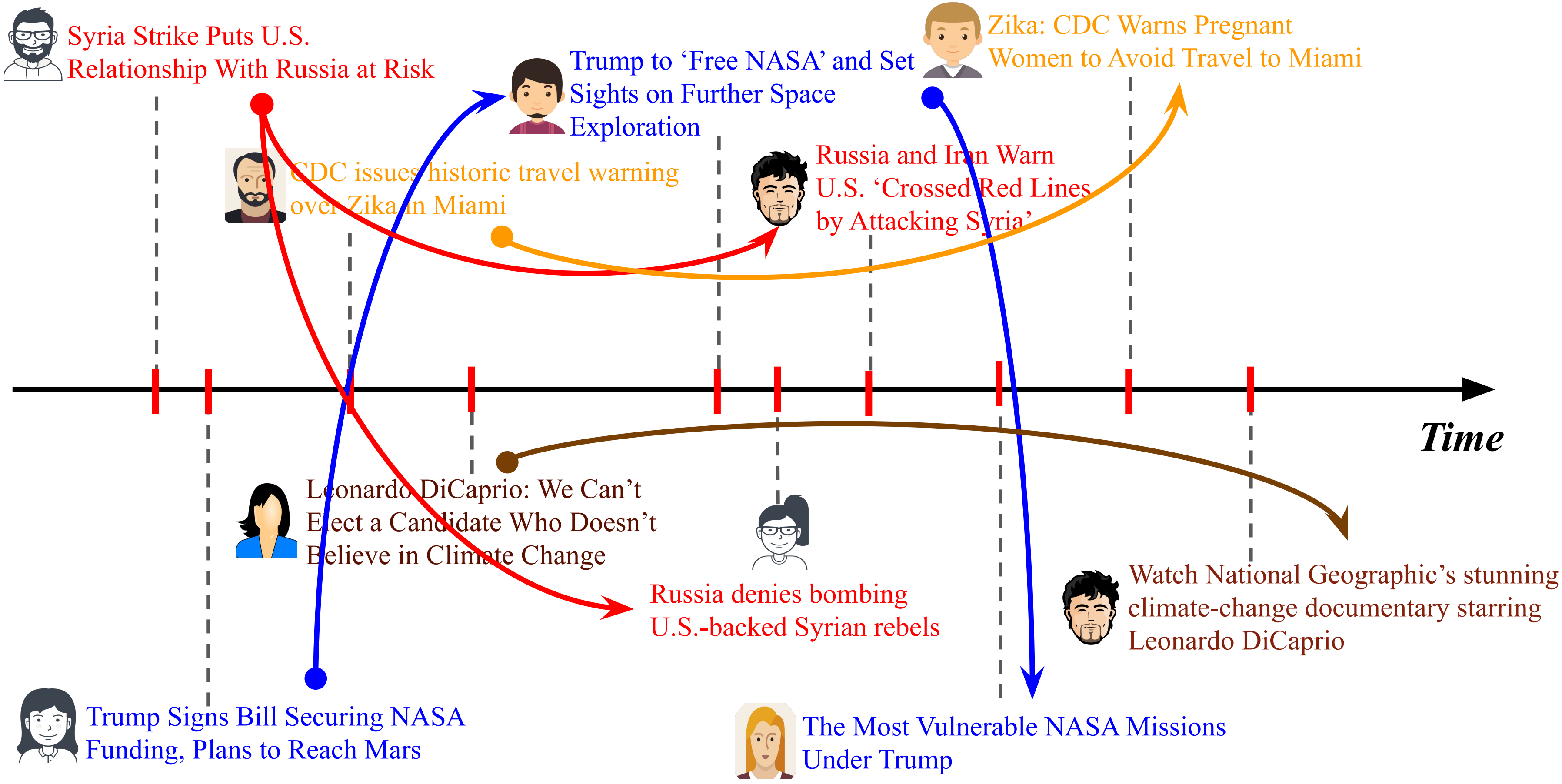
- User Temporal Dynamics
- Preferred topics of each user
- Network Strengths (user-user influence)

- Topics
- Topical Interactions

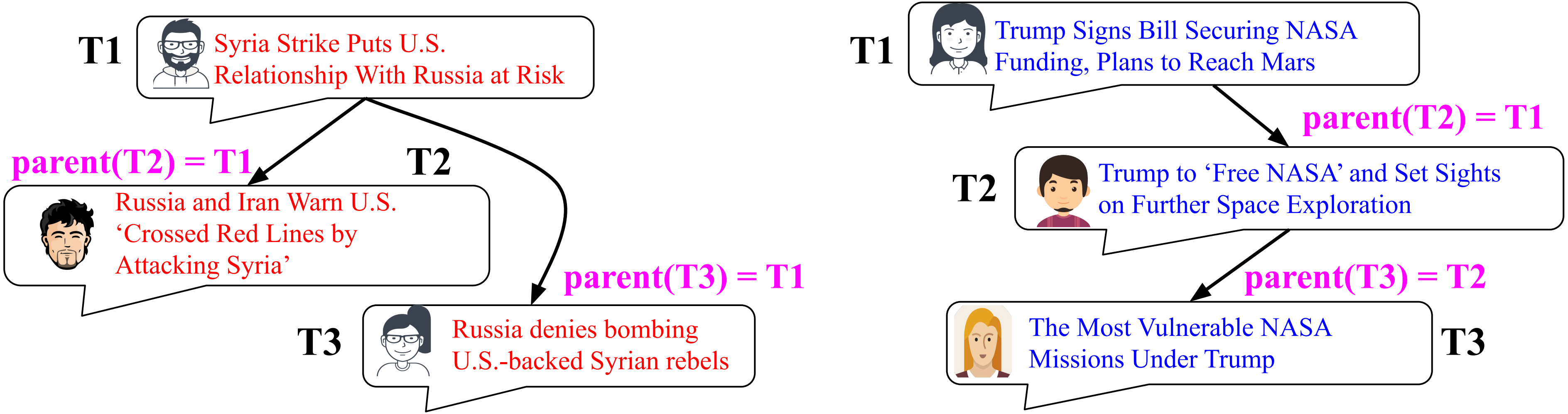
Data: Network + Time-series of Tweets



Mixture of Conversations



Cascades (Separate Conversations)



Separate these conversations out!!!

Hidden Markov Hawkes Process

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

- Coupled inference: Collapsed Gibbs sampling

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

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

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 $x_{e,w} \sim Discrete_{\mathcal{W}}(\boldsymbol{\zeta}_{\eta_e})$

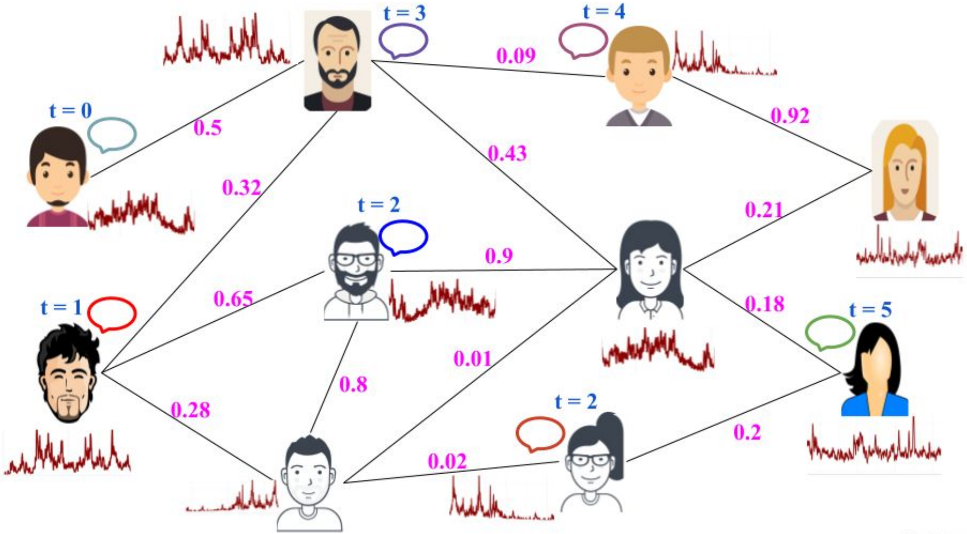
Temporal Dynamics and Network Inference using Multivariate Hawkes Process

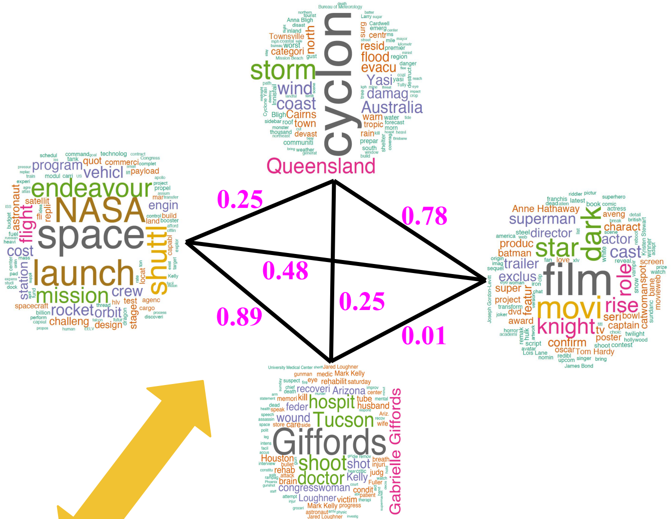
Cascade reconstruction and Topical Interactions coupling Multivariate Hawkes Process and Topical Markov Chains

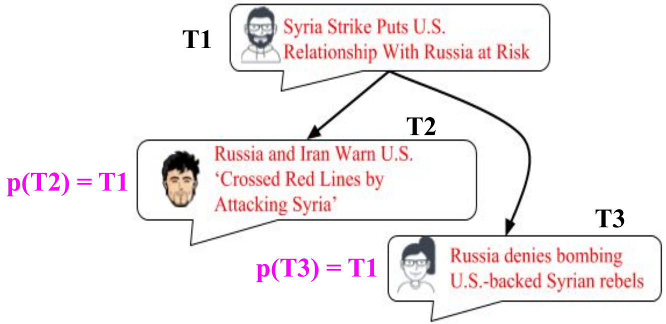
Topic Model

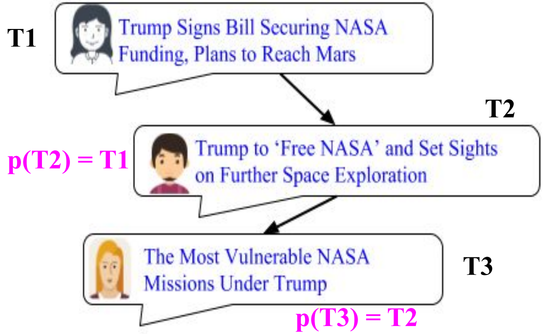
Inference

Challenge - Coupled Problems



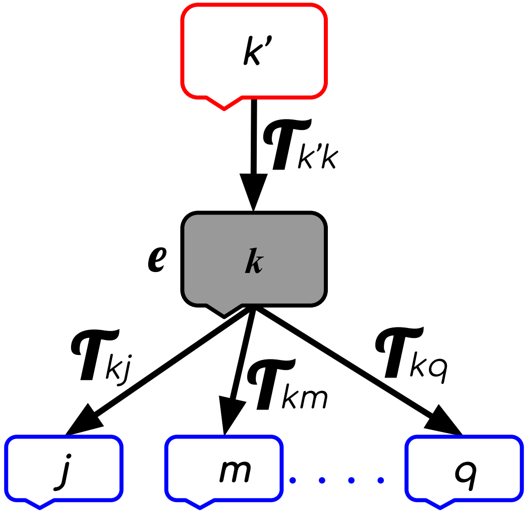






Topic Inference

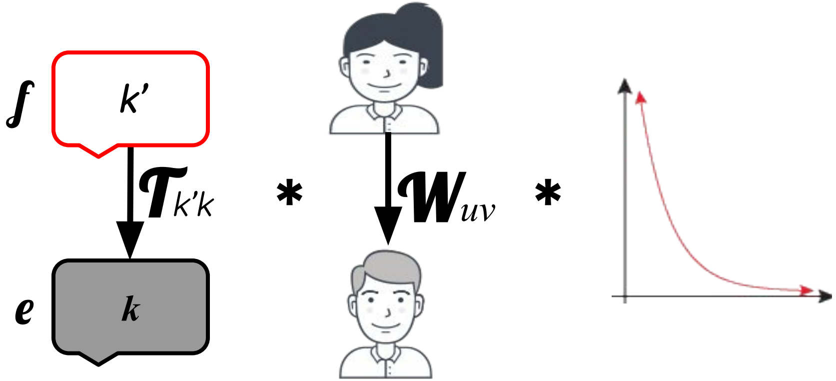
$$\mathcal{P}\left(\mathit{topic}(e) = k \,\middle|\, \mathit{parentStructure}, \mathit{tweetText}, \{\mathit{topic}(f) \mid f \mathrel{!=} e\} \right) \propto$$



The Most Vulnerable NASA Missions Under Trump

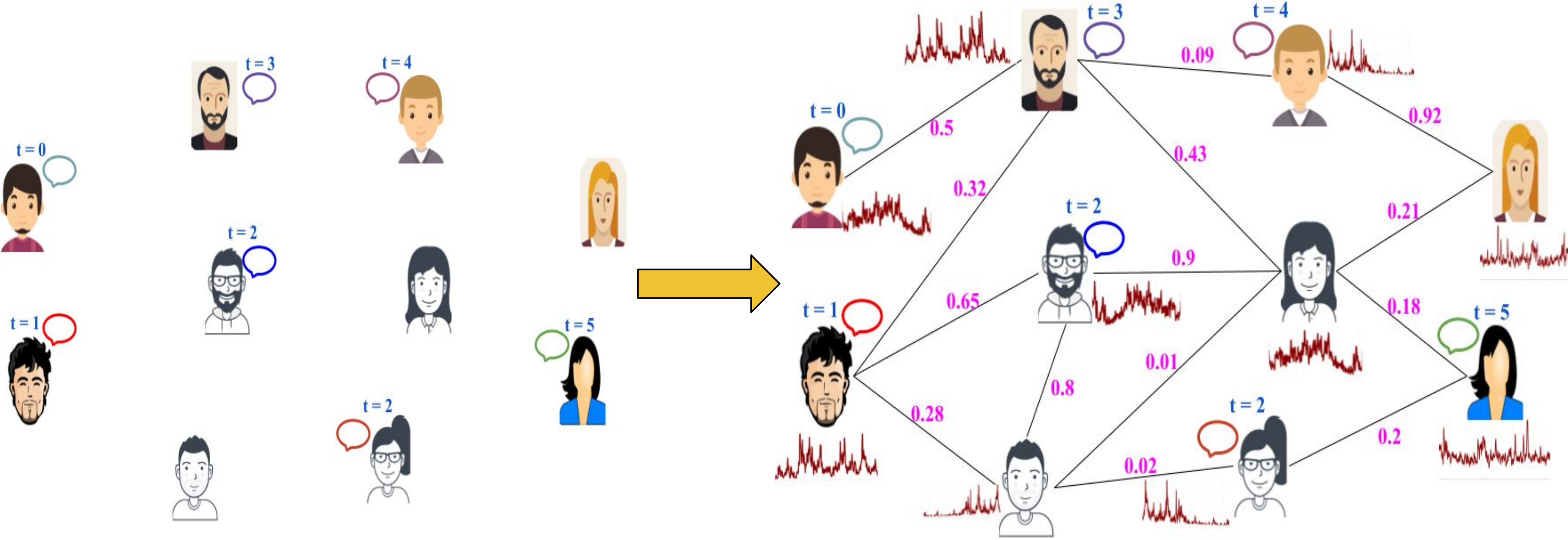
Cascade Inference

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



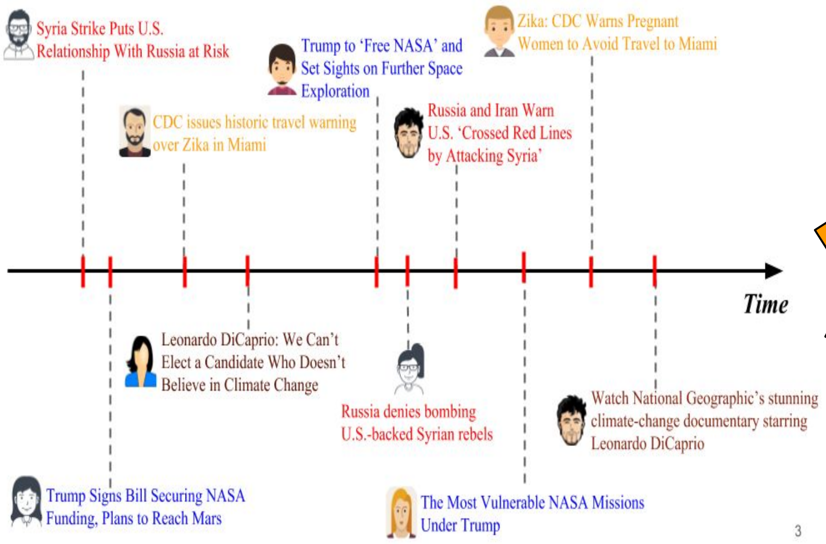
Existing Models

Network Hawkes Model

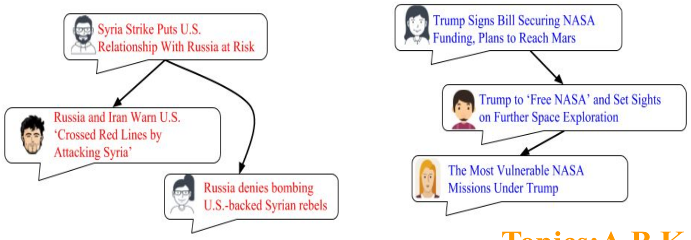


Does not model (textual) content of events / tweets

Hawkes Topic Model (HTM) [He et al. '15]

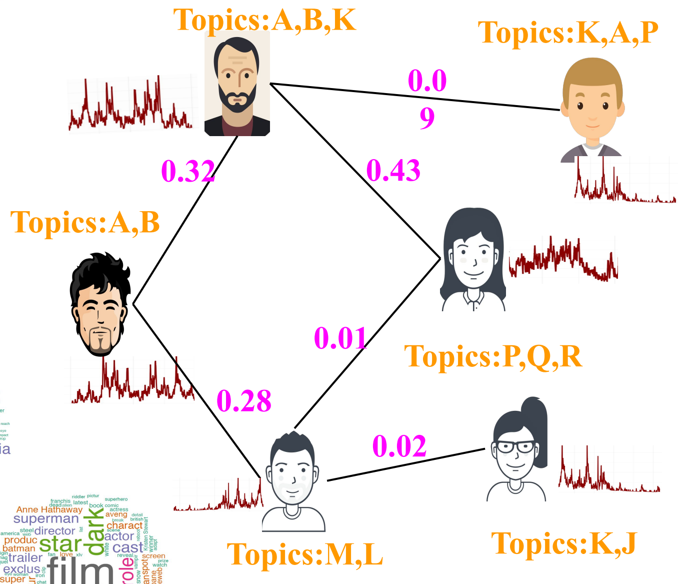


Set of tweets and time-stamps



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

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

v/s

Generation of Topic of child event in HMHP

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

$$Topic(e) \sim \zeta(Topic(parent(e)))$$

where, ζ 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

	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

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

Network Reconstruction Error

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

Cascade Reconstruction Accuracy

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

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

Comparison with HTM [He et al. '15]

Synthetic events generated using HMHP model

Window Length	1000	2000	3000	4000	5000
HTM	2.811	1.982	1.464	1.292	1.351
HMHP	1.297	0.925	0.677	0.646	0.657

Network Inference (TAE) ***(lower the better)***

Window Length	1000	2000	3000	4000	5000
HTM	0.681	0.687	0.712	0.716	0.708
HMHP	0.926	0.924	0.95	0.94	0.935

Parent Identification (Accuracy) ***(higher the better)***

Comparison with HTM [He et al. '15]

Synthetic events (short documents) generated using HTM model

Window Length	1000	2000	3000	4000	5000
HTM	3.167	2.377	2.014	1.964	1.519
HMHP	1.696	1.200	1.168	1.396	1.243

Network Inference (TAE) ***(lower the better)***

Window Length	1000	2000	3000	4000	5000
HTM	0.575	0.588	0.61	0.618	0.628
HMHP	0.716	0.730	0.736	0.730	0.748

Parent Identification (Accuracy) ***(higher the better)***

Generative Model

What all to model?

- Temporal Dynamics for each user
- User Network Strengths
- Topics
- Topical Interactions
- Topic preference for each user

Modeling Time + Network: Hawkes Process

$$\lambda_v(t) = \mu_v(t) + \sum_{n=1}^{|\mathcal{H}_{t-}|} h_{c_n,v}(t - t_n)$$

Modeling Time + Network: Multivariate Hawkes Process

$$\lambda_v(t) = \underbrace{\boxed{\mu_v(t)}}_{\text{Base Intensity}} + \underbrace{\boxed{\sum_{n=1}^{|\mathcal{H}_{t-}|} h_{c_n,v}(t - t_n)}}_{\text{Excitation because of History (Neighbors)}}$$

$$h_{u,v}(\Delta t) = W_{u,v} f(\Delta t)$$

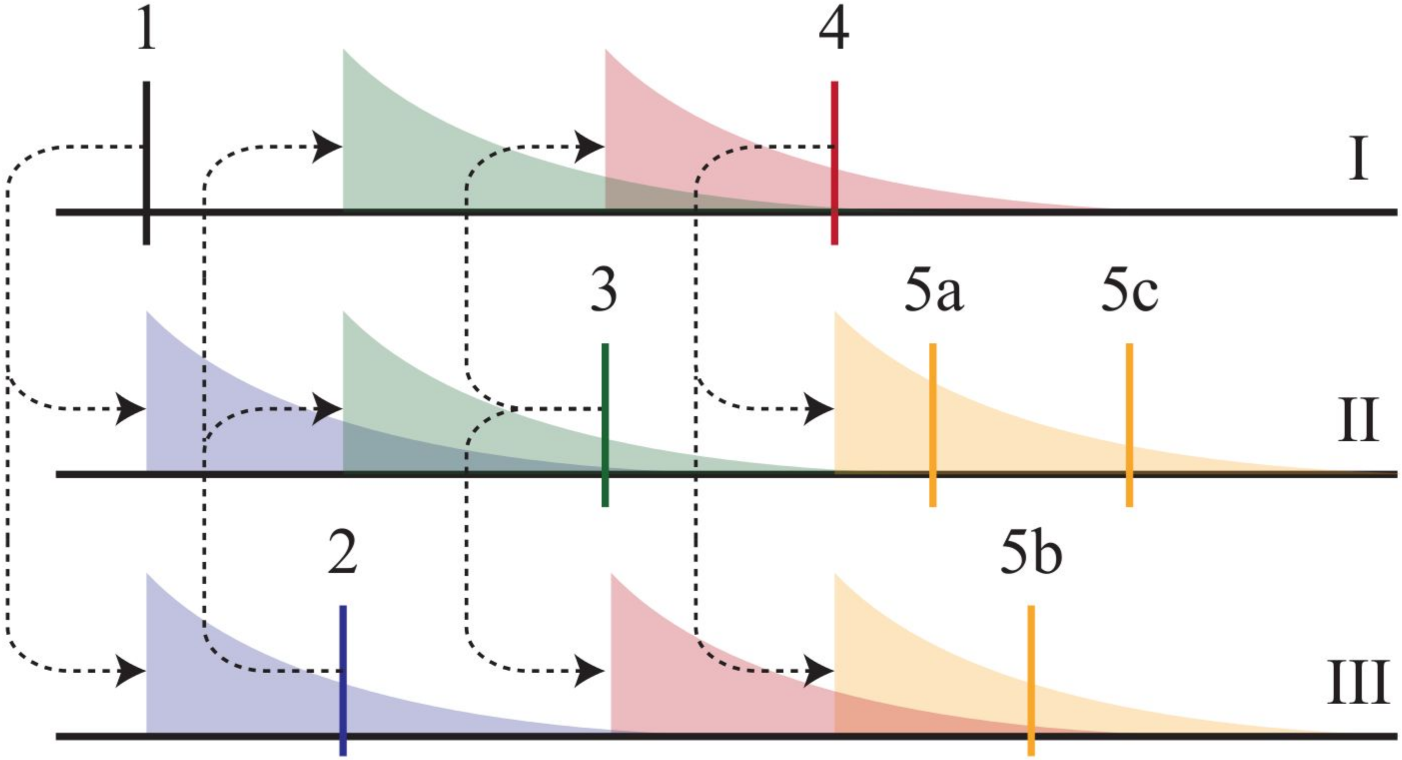
Level wise event generation [A. Simma 2010]

- Draw (spontaneous) events for each user with the base intensity -- *(Level-0 events)*.
- Subsequent events are drawn using the following non-homogenous Poisson process

$$\Pi_l \sim Poisson\left(\sum_{(t_n, c_n, z_n) \in \Pi_{l-1}} h_{c_n, \cdot}(t, t_n)\right)$$

Level-i event can be anywhere on the timeline, it's just that the timestamps of level-i events is greater than the timestamps of level-(i-1) events

Modeling Time + Network: Multivariate Hawkes Process



HMHP Generative Process

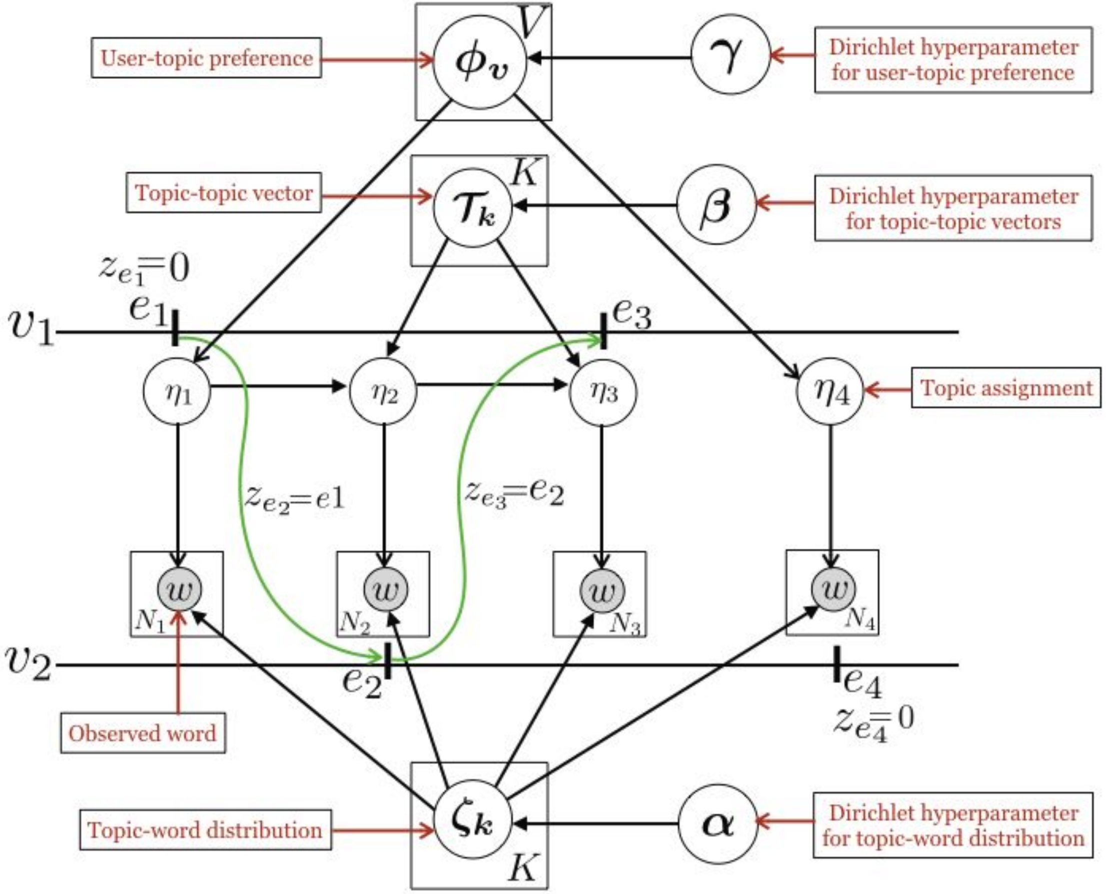
1) Generate (t_e, c_e, z_e) for all events according Multivariate Hawkes Process.
2) For each topic k: sample $\boldsymbol{\zeta}_k \sim Dir_{\mathcal{W}}(\boldsymbol{\alpha})$
3) For each topic k: sample $\boldsymbol{\mathcal{T}}_k \sim Dir_K(\boldsymbol{\beta})$
4) For each node v: sample $\boldsymbol{\phi}_v \sim Dir_K(\boldsymbol{\gamma})$
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

Generative Model



Inference

Joint Probability

$$P(E, \boldsymbol{\Phi}, \boldsymbol{\mathcal{T}}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{W}, \boldsymbol{\mu}) =$$

$$\prod_{v \in V} P(\boldsymbol{\phi}_v \mid \boldsymbol{\gamma}) \times \prod_{k=1}^{K} P(\boldsymbol{\zeta}_k \mid \boldsymbol{\alpha}) \times \prod_{k=1}^{K} P(\boldsymbol{\mathcal{T}}_k \mid \boldsymbol{\beta})$$

$$\times \prod_{e \in E} \left\{ \left[\prod_{e': t_{e'} < t_e} P(\eta_e | \boldsymbol{\mathcal{T}}_{\eta_{z_e}})^{\delta_{z_e, e'}} \right] P(\eta_e | \boldsymbol{\phi}_v)^{\delta_{z_e, 0}} \right\}$$

$$\times \prod_{e \in E} \left[\prod_{w=1}^{N_e} P(x_{e,w} | \eta_e, \boldsymbol{\zeta}_{\eta_e}) \right]$$

$$\times \prod_{v \in V} \left[exp\left(-\int_0^T \mu_v(\tau) d\tau \right) \prod_{e \in E} \mu_v(t_e)^{\delta_{c_e, v} \delta_{z_e, 0}} \right]$$

$$\times \prod_{e \in E} \prod_{v \in V} \left[exp\left(-\int_{t_e}^T h_{c_e, v}(\Delta\tau) d\tau \right) \prod_{e' \in E} h_{c_e, c_{e'}}(\Delta(t_{e'}))^{\delta_{c_{e'}, v} \delta_{z_{e'}, e}} \right]$$

Joint Probability

$$P(E, \boldsymbol{\Phi}, \boldsymbol{\mathcal{T}}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{W}, \boldsymbol{\mu}) =$$

$$\prod_{v \in V} P(\boldsymbol{\phi}_v \mid \boldsymbol{\gamma}) \times \prod_{k=1}^{K} P(\boldsymbol{\zeta}_k \mid \boldsymbol{\alpha}) \times \prod_{k=1}^{K} P(\boldsymbol{\mathcal{T}}_k \mid \boldsymbol{\beta})$$

- Priors

$$\times \prod_{e \in E} \left\{ \left[\prod_{e': t_{e'} < t_e} P(\eta_e | \boldsymbol{\mathcal{T}}_{\eta_{z_e}})^{\delta_{z_e, e'}} \middle| P(\eta_e | \boldsymbol{\phi}_v)^{\delta_{z_e, 0}} \right] \right\}$$

- Topic Transitions/Interactions
- Topics for spontaneous events

$$\times \prod_{e \in E} \left[\prod_{w=1}^{N_e} P(x_{e,w} | \eta_e, \boldsymbol{\zeta}_{\eta_e}) \right]$$

- Generating words for each doc

$$\times \prod_{v \in V} \left[exp\left(- \int_0^T \mu_v(\tau) d\tau \right) \prod_{e \in E} \mu_v(t_e)^{\delta_{c_e, v} \delta_{z_e, 0}} \right]$$

- Time (spontaneous svents)

$$\times \prod_{e \in E} \prod_{v \in V} \left[exp\left(- \int_{t_e}^T h_{c_e, v}(\Delta\tau) d\tau \right) \prod_{e' \in E} h_{c_e, c_{e'}}(\Delta(t_{e'}))^{\delta_{c_{e'}, v} \delta_{z_{e'}, e}} \right]$$

- Time (Influenced Events)

Topic Inference

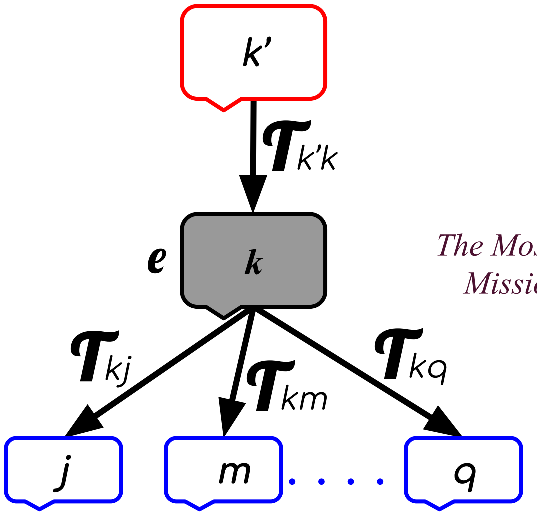
$$\mathcal{P}\left(\begin{matrix}\textit{topic}(e) = k \mid \textit{parentStructure}, \textit{tweetText}, \\ \{\textit{topic}(f) \mid f \mathrel{!=} e\}\end{matrix}\right) \propto \frac{\beta_k + N_{k',k}^{(\neg(z_e,e))}}{(\sum_l \beta_l) + \boldsymbol{N}_{k'}^{(\neg(z_e,e))}} \times \frac{\prod_{w\in d_e}\prod_{i=0}^{N_e^w-1}(\alpha_w + \mathfrak{T}_{k,w}^{\neg e} + i)}{\prod_{i=0}^{N_e-1}((\sum_{w\in\mathcal{W}}\alpha_w) + \mathfrak{T}_k^{\neg e} + i)}$$

$$\times \frac{\prod_{l'=1}^{K}\prod_{i=0}^{N_{k,l'}^{(C_e)}-1}(\beta_{l'} + N_{k,l'}^{(\neg C_e)} + i)}{\prod_{i=0}^{\boldsymbol{N}_k^{(C_e)}-1}((\sum_{l'}\beta_{l'}) + \boldsymbol{N}_k^{\neg C_e} + i)}$$

Topic Inference

$$\mathcal{P}\left(\begin{array}{c} topic(e) = k \mid parentStructure, tweetText, \\ \{topic(f) \mid f \mathrel{!=} e\} \end{array}\right) \propto \frac{\beta_k + N_{k',k}^{(\neg(z_e,e))}}{(\sum_l \beta_l) + \boldsymbol{N}_{k'}^{(\neg(z_e,e))}} \times \frac{\prod_{w\in d_e}\prod_{i=0}^{N_e^w-1}(\alpha_w + \mathfrak{T}_{k,w}^{\neg e}+i)}{\prod_{i=0}^{N_e-1}((\sum_{w\in\mathcal{W}}\alpha_w)+\mathfrak{T}_k^{\neg e}+i)} \times \frac{\prod_{l'=1}^{K} \prod_{i=0}^{N_{k,l'}^{(C_e)}-1}(\beta_{l'}+N_{k,l'}^{(\neg C_e)}+i)}{\prod_{i=0}^{\boldsymbol{N}_k^{(C_e)}-1}((\sum_{l'}\beta_{l'})+\boldsymbol{N}_k^{\neg C_e}+i)}$$

$$\mathcal{P}\left(\begin{array}{c} topic(e) = k \mid parentStructure, tweetText, \\ \{topic(f) \mid f \mathrel{!=} e\} \end{array}\right) \propto$$



The Most Vulnerable NASA Missions Under Trump

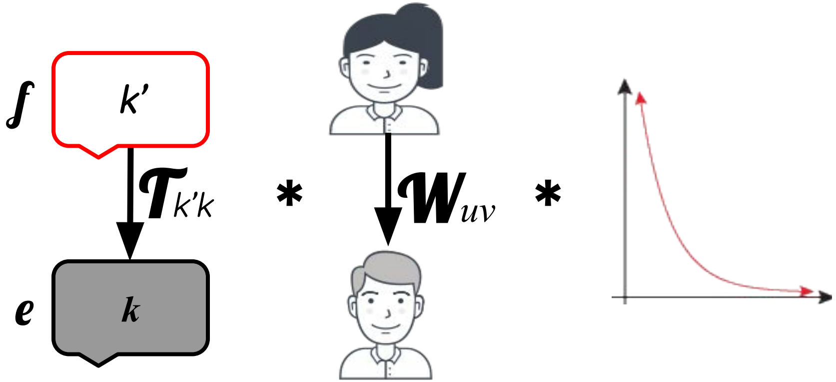
Cascade Inference

$$\mathcal{P}\left[par(e) = f \mid Topics, W, \mu, timeStamps\right] \propto \frac{(\beta_k + N_{k',k} - 1)}{((\sum_{k=1}^{K} \beta_k) + \boldsymbol{N}_{k'} - 1)} \times h_{u_{e'},u_e}(t_e - t_{e'})$$

Cascade Inference

$$\mathcal{P}\left[par(e) = f \mid Topics, W, \mu, timeStamps\right] \propto \frac{(\beta_k + N_{k',k} - 1)}{((\sum_{k=1}^{K} \beta_k) + \mathbf{N}_{k'} - 1)} \times h_{u_{e'},u_e}(t_e - t_{e'})$$

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



Network Inference

$$P(W_{u,v} = x \mid E_t^{(u,v)}, \boldsymbol{z}) \propto \quad x^{\alpha_1} \exp(-x\beta_1)$$

where,

$$\alpha_1 = (N_{u,v} + \alpha - 1)$$

$$\beta_1 = (N_u + \frac{1}{\beta})^{-1}$$

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*

Knowledge in HMHP

Topical Structure

Tonight, we heard from two candidates -- but only one president. #ImWithHer

8:13 AM - 27 Sep 2016

7,060 Retweets **20,150** Likes

"Hillary won big time. It was a shut out." -- @HardballChris #debatenight

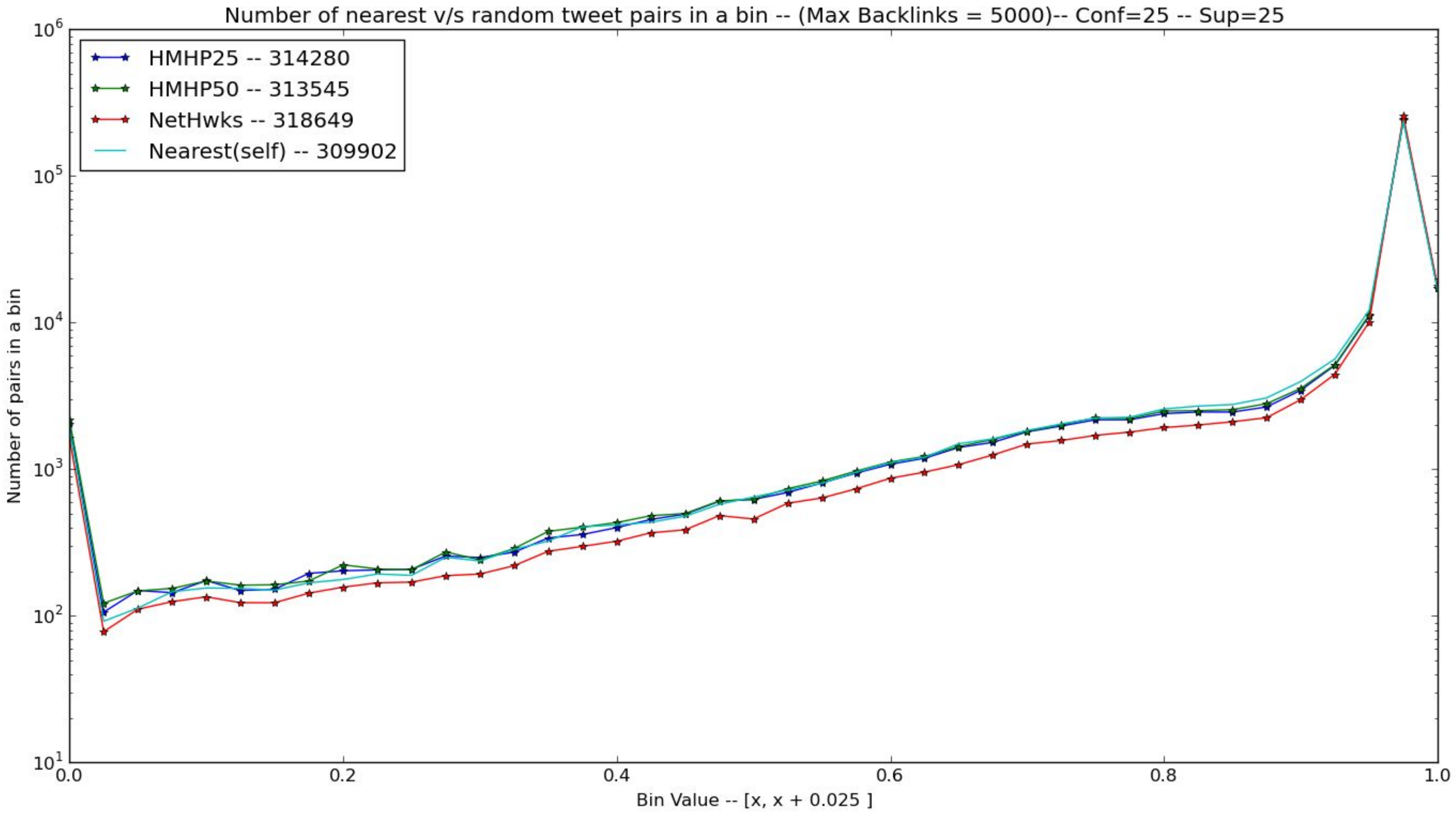
8:14 AM - 27 Sep 2016

4,313 Retweets **12,374** Likes

Topical Structure

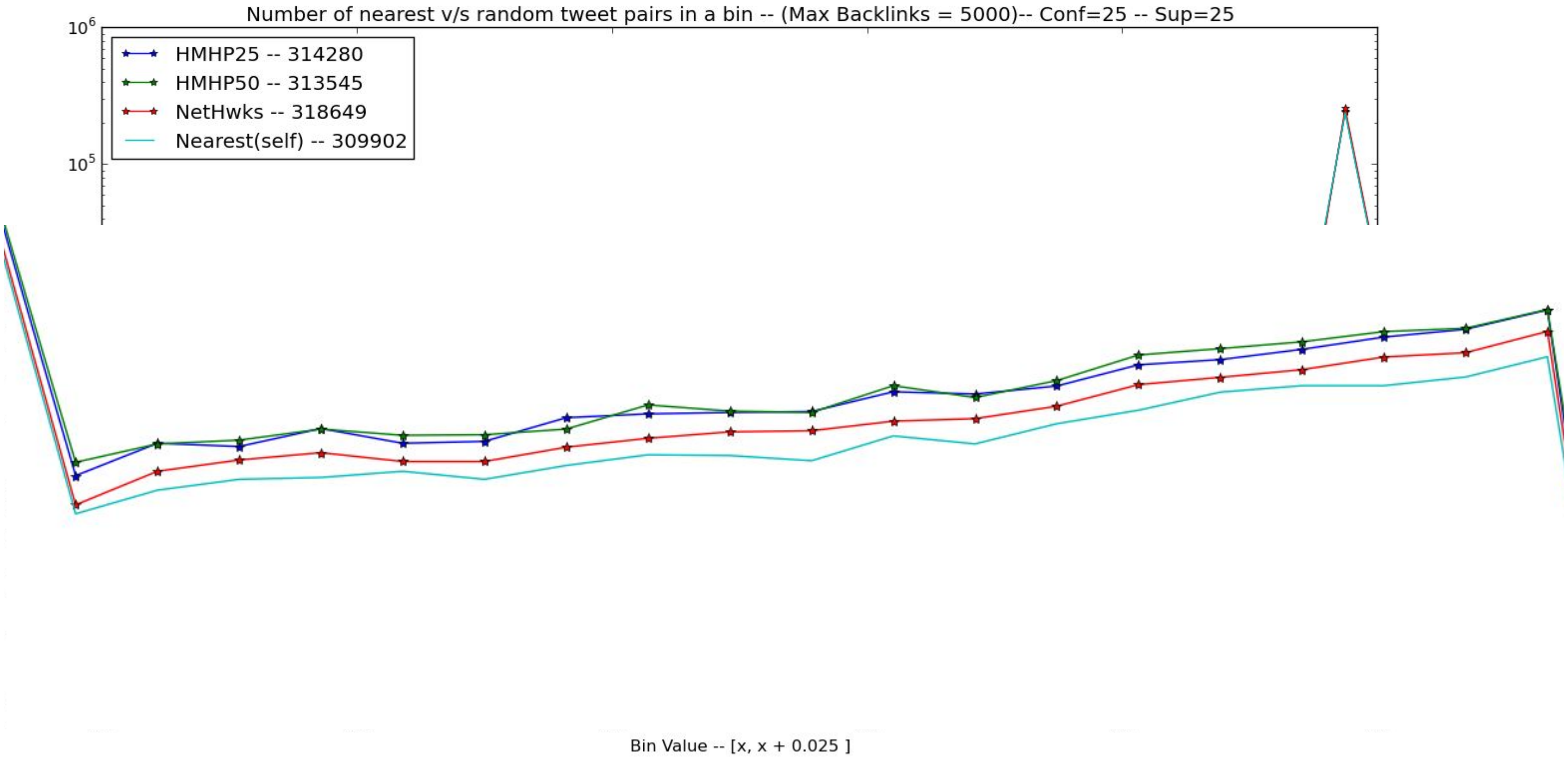
Entities in parent-child tweet pairs are "closer" on Wiki?

Jaccard Distance between Parent-Child Tweets



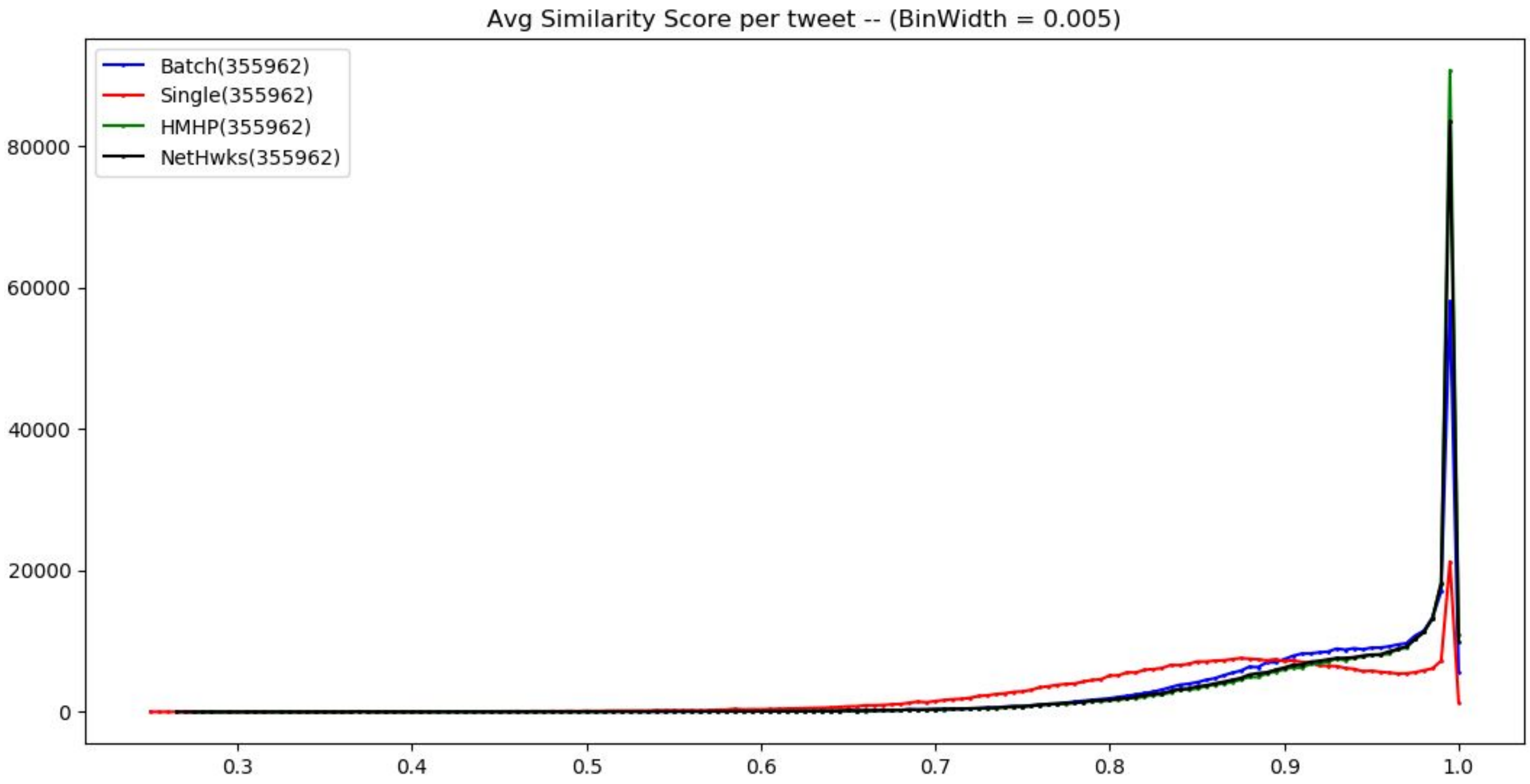
Note: Annotation is done using DBPedia Spotlight

Jaccard Distance between Parent-Child Tweets

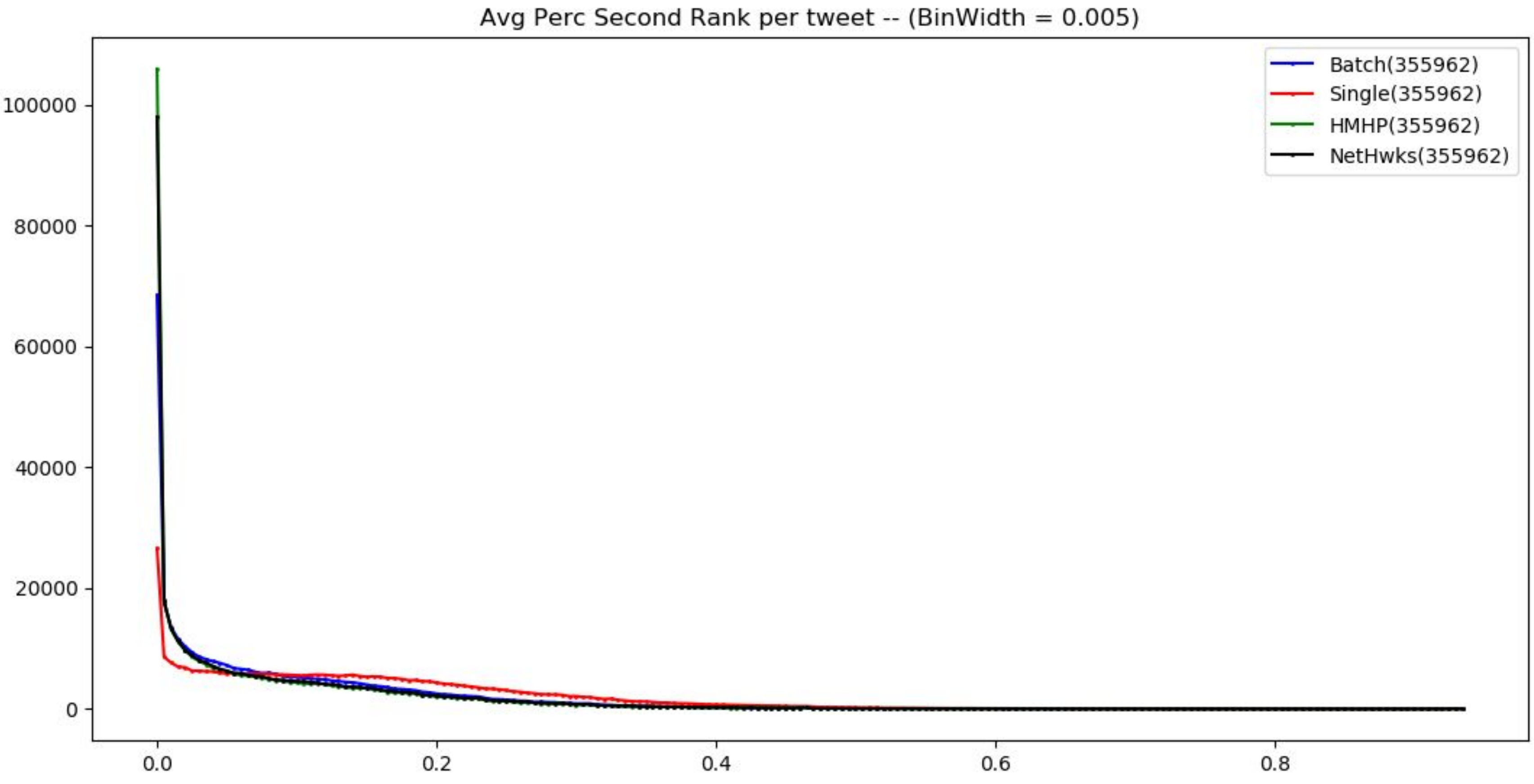


Better Cascades Better Annotation?

Avg. Similarity Score for Cascades



Avg. Percentage Second Rank for Cascades



Coupling Cascades and Entity Identification

- *Better parent-child identification (cascade construction) can help in better annotation (entity identification)*

- *Better annotation can help in better parent-child identification (cascade construction)?*

Goal

A Generative model for textual time-series data from user networks having topical interactions along with structure among topics

Thank You