Models and Algorithms for Information Diffusion

Jayesh Choudhari

Under the guidance of: *Prof. Anirban Dasgupta*

Mentor: *Dr. Indrajit Bhattacharya (TCS Research)*

DSC Committee:

Prof. Neeldhara Misra, Prof. Chetan Pahlajani

Overview

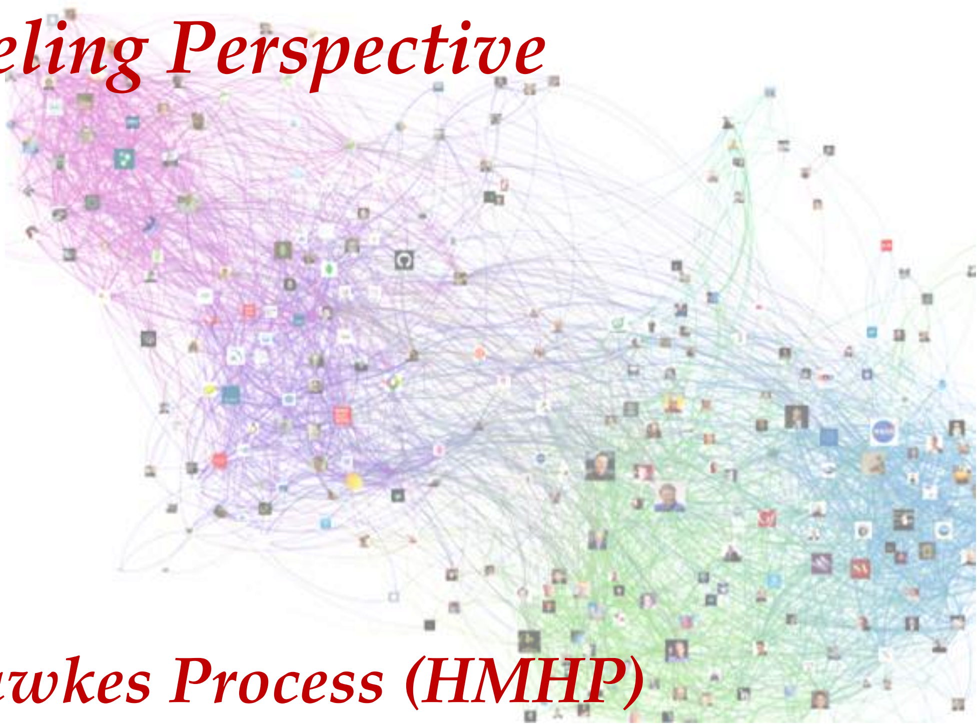
Part-I Modeling Perspective

- *Discovering topical interactions in text-based cascades using* ***Hidden Markov Hawkes Processes (HMHP)****.*
  Choudhari, J.*, Dasgupta, A., Bhattacharya, I., & Bedathur, S. (2018, November). In 2018 IEEE International Conference on Data Mining (ICDM)*
- *Unified Marked Temporal Point Process. (****Dual Network Hawkes Process (DNHP)****)*
  Choudhari, J.*, Dasgupta, A., Bhattacharya, I., & Bedathur, S. In* ***Temporal Point Process Workshop – NeurIPS 2019****.*

Part-II Algorithmic Perspective

- *Saving Critical Nodes with* ***Firefighters*** *is FPT.* ***Choudhari, J.****, Dasgupta, A., Misra, N., & Ramanujan, M. S. (2017). In 44th International Colloquium on Automata, Languages, and Programming* ***(ICALP 2017)****.*

Part-I: Modeling Perspective



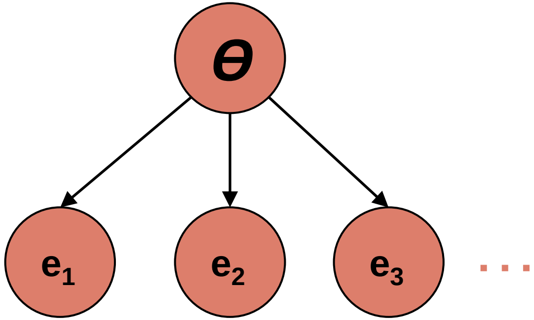
Hidden Markov Hawkes Process (HMHP)
&
Dual Network Hawkes Process (DNHP)

What is a Model?

A model is an imaginary procedure that generates observed data e_1, e_2, ...

The procedure has a parameter(s) ***θ****, and we use data to learn it*

"inference"



Why should we learn the parameter(s) θ?

Lets us ***make predictions*** *and* ***answer questions*** *about data at an abstract, high level*

Example

T H T H H H

From the data directly, one can answer simple questions:

"How many heads/tails?"

"Was there a heads before a tails?"

"What was the longest string of heads?"

Example

T H T H H H

How about using a **model?**:

Every coin toss has a probability θ of being heads; estimate $\theta \approx 4/(4+2) = 2/3$

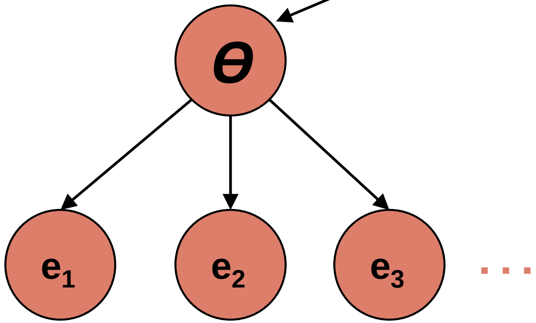
Now we can make predictions and ask abstract questions:

"How likely am I to see heads next?"

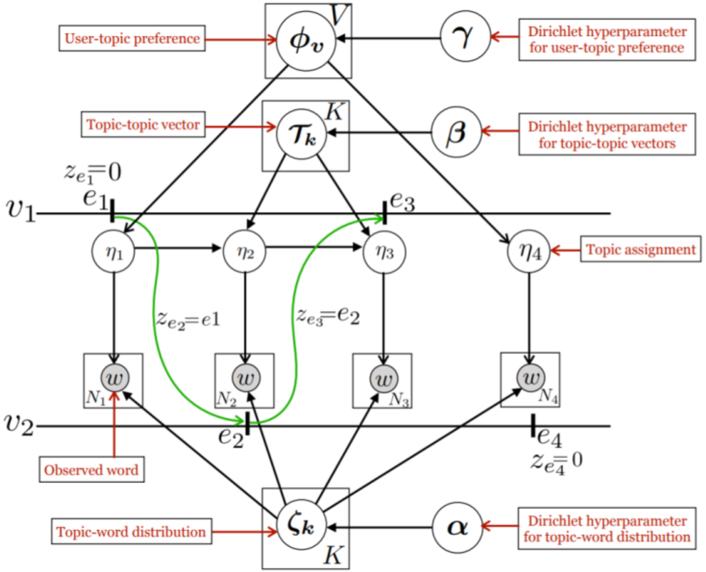
"Is this coin fair?"

Bayesian Models

random parameter (unknown)



data is observed/fixed



Bayesian models help in ***quantifying uncertainty*** in the unknown parameter(s) $\boldsymbol{\theta}$

Learn distribution over the values that the parameter(s) can take

probably not very certain that $\boldsymbol{\theta} = 2/3$

Bayesian Models - Applications

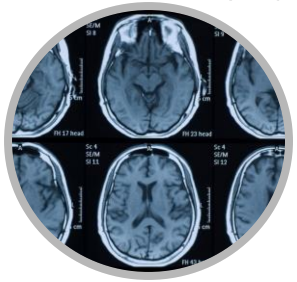
suggesting movies

predicting crop yields

driving autonomously

summarizing topics

analyzing medical imagery



making smart trades

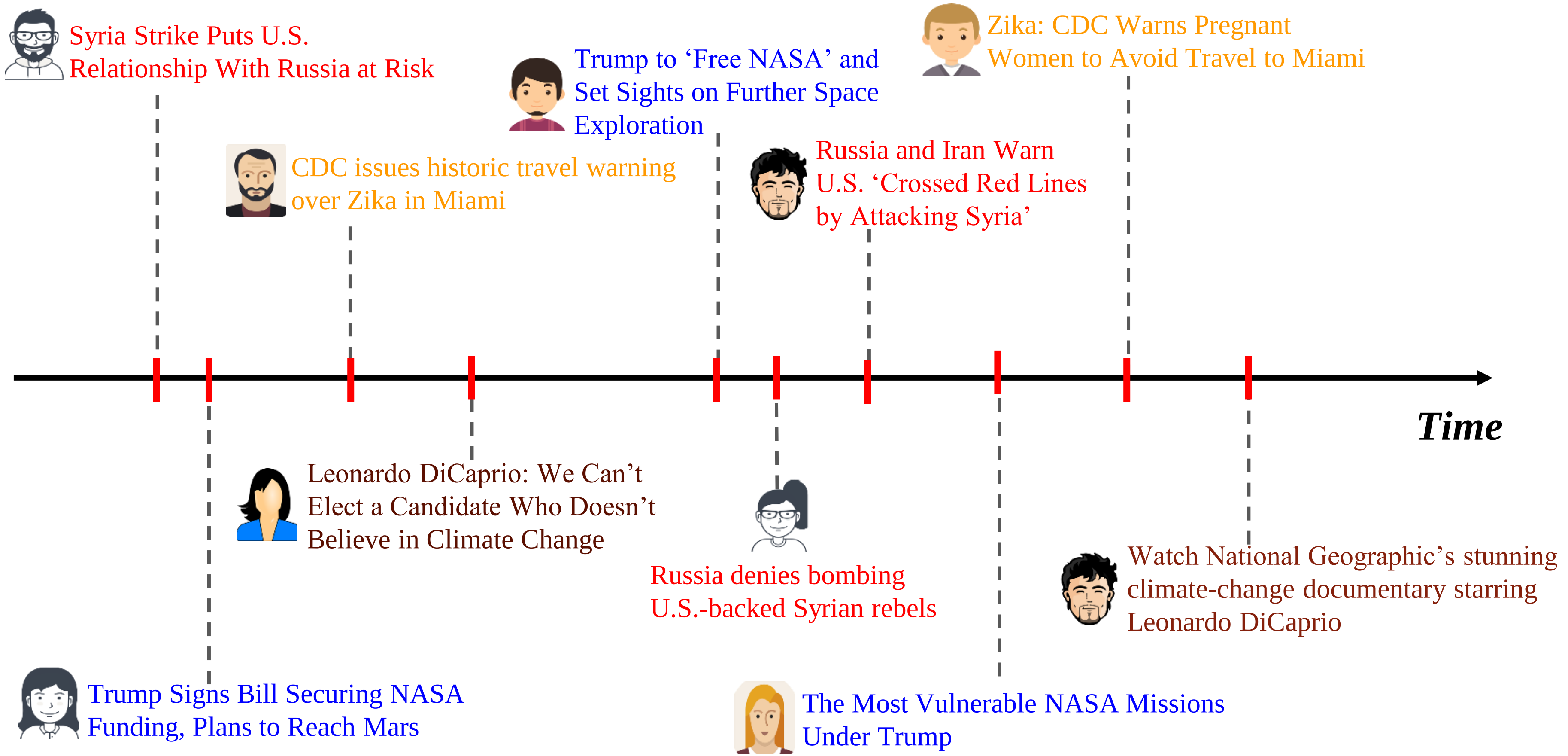
discovering protein structure

exploring the universe

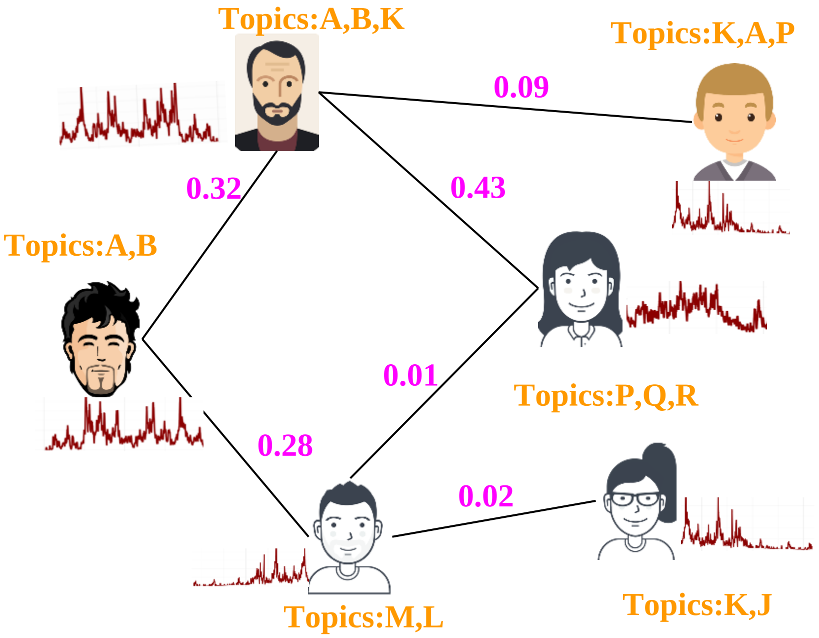
reducing traffic

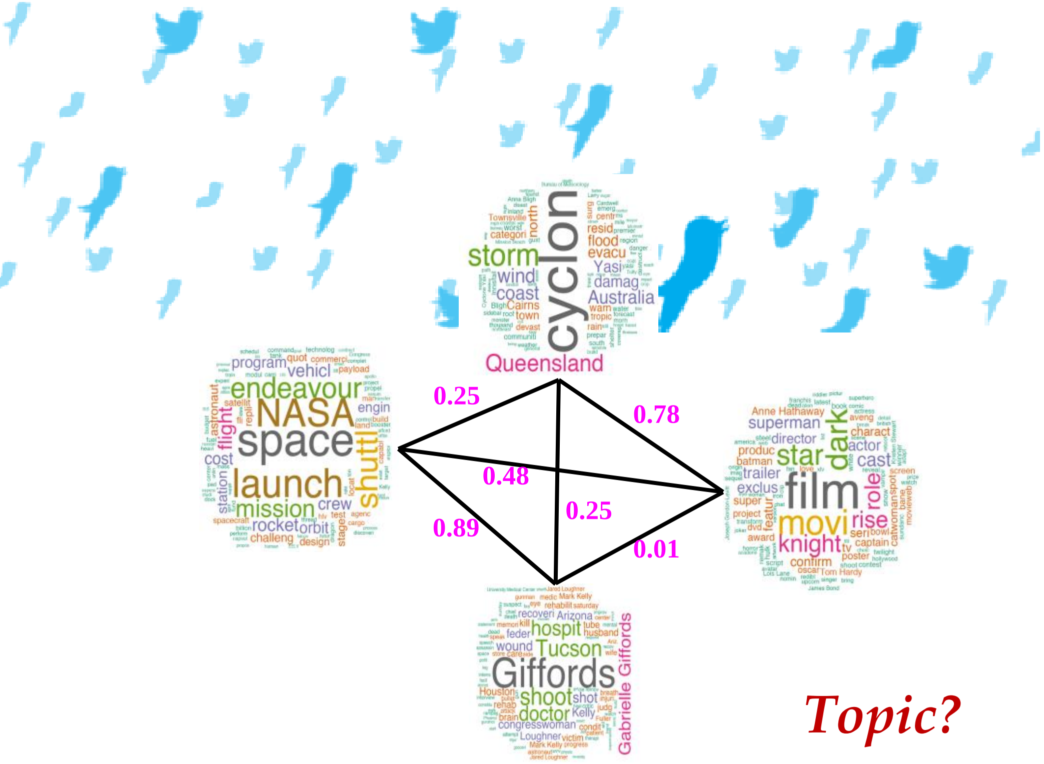
spread of ‘information’

Data: Network + Time Series of Tweets



What is there to do?





User Temporal Dynamics

Preferred topics of each user

Network Strengths (user-user influence)

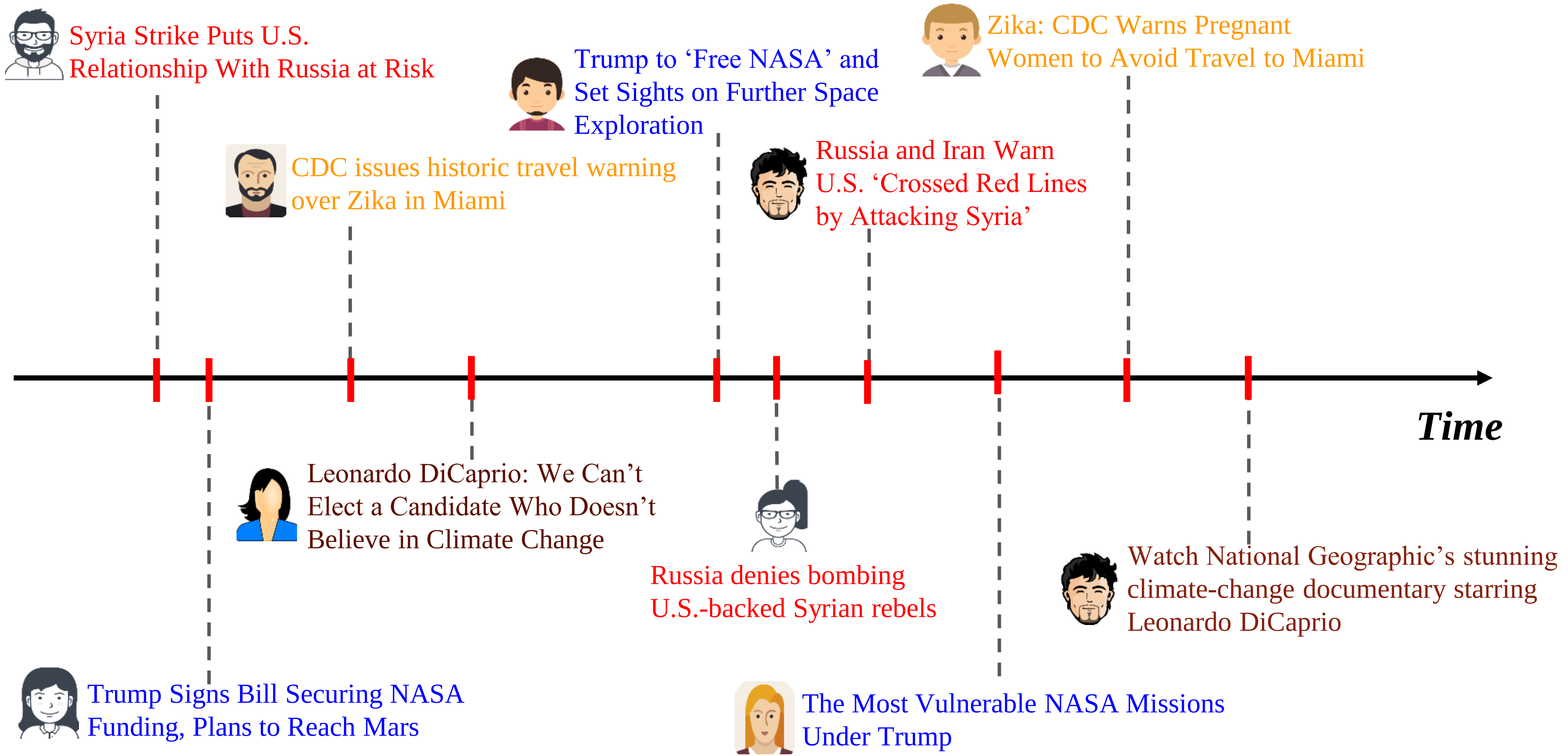
Topics in the tweets/documents

Topical Interactions

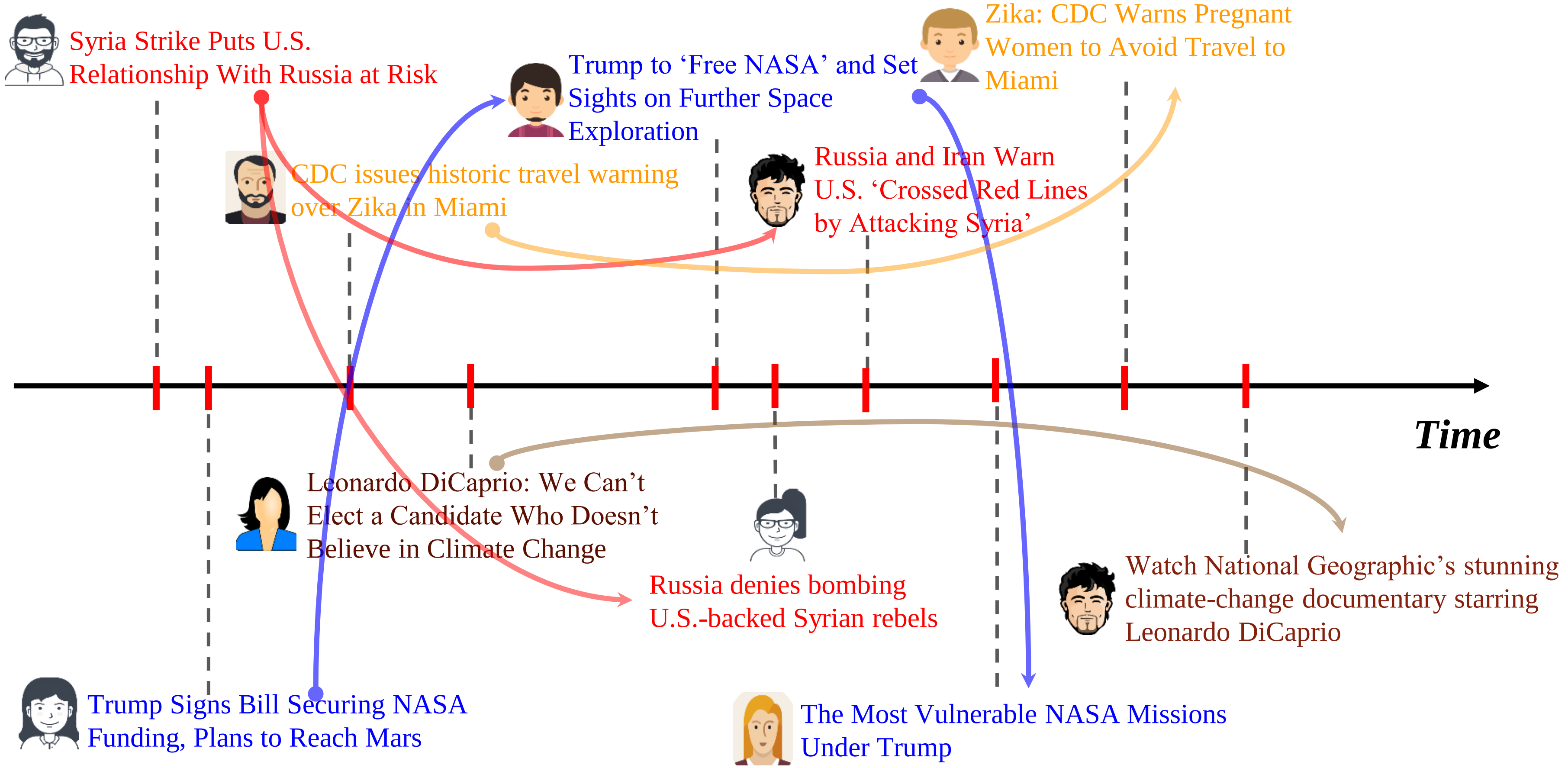
Research Gaps + **Gaps Filled**

User N/w	*Time-series*	*Topic*	*Topic N/w*	*Topic excitations*	*Past Work (+ Thesis)*
✗	✓	✓	✗	✓	*Du et al. ['15] (Dirichlet-Hawkes)*
✓	✓	✗	✗	✗	*Simma-Jordan ['10],* *Gomez et al. ['11] (NetInf, NetRate)* *Yang et al. ['13] (MMHP),* *Linderman et al. ['14] (NetHawkes)*
✓	✓	✓	✗	✗	*He et al. ['15] (HTM)*
✓	✓	✓	✓	✗	***Choudhari et al. ['18] (HMHP)***
✓	✓	✓	✓	✓	***Choudhari et al. ['19] (DNHP)***

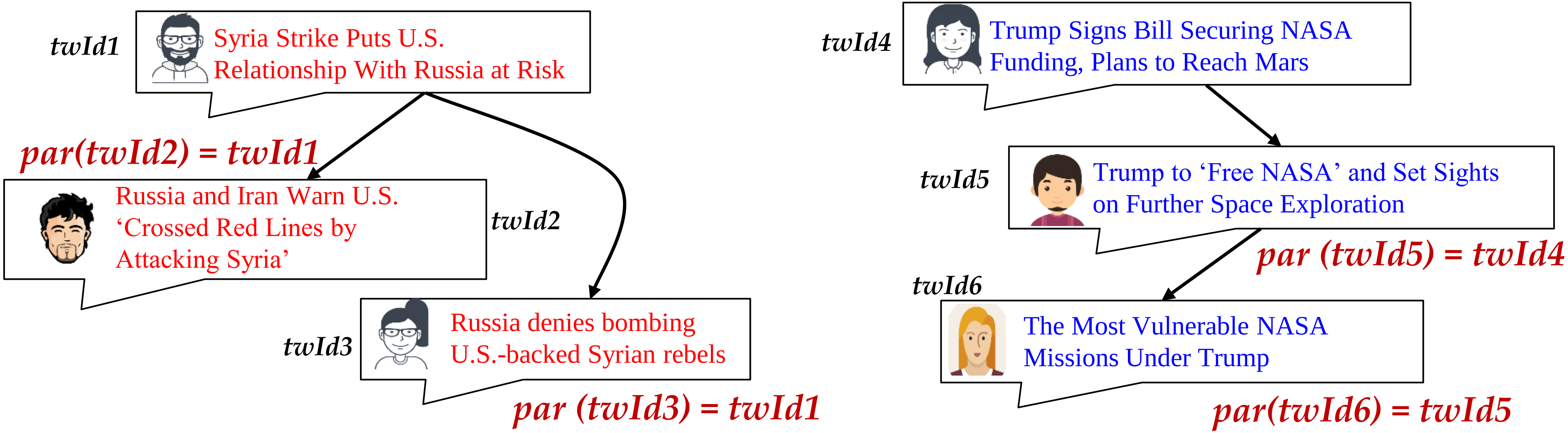
Data: Network + Time Series of Tweets



Mixture of Conversations



Cascades: Separate Conversations



Just separate these conversations!!!

User Temporal Dynamics

Network Strengths (user-user influence)

Preferred topics of each user

Topics in the tweets/documents

Topical Interactions

Why Topical Interactions in HMHP?

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

HMHP *(HTM)* Generative model: Overview

1. Generate time-stamps of events for each user (Hawkes Process)

2. For each event, assign topic (dependent on the topic of the influencer event i.e. parent event)

3. For each event, generate words

Self-exciting Point Process (Hawkes Process)

Time-stamps are characterized by an intensity function:

$$\lambda(t)dt := Pr\,(event\ in\ [t + dt) \,|\, \mathcal{H}_{t^-})$$

Time Kernel

Multivariate Hawkes Process

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

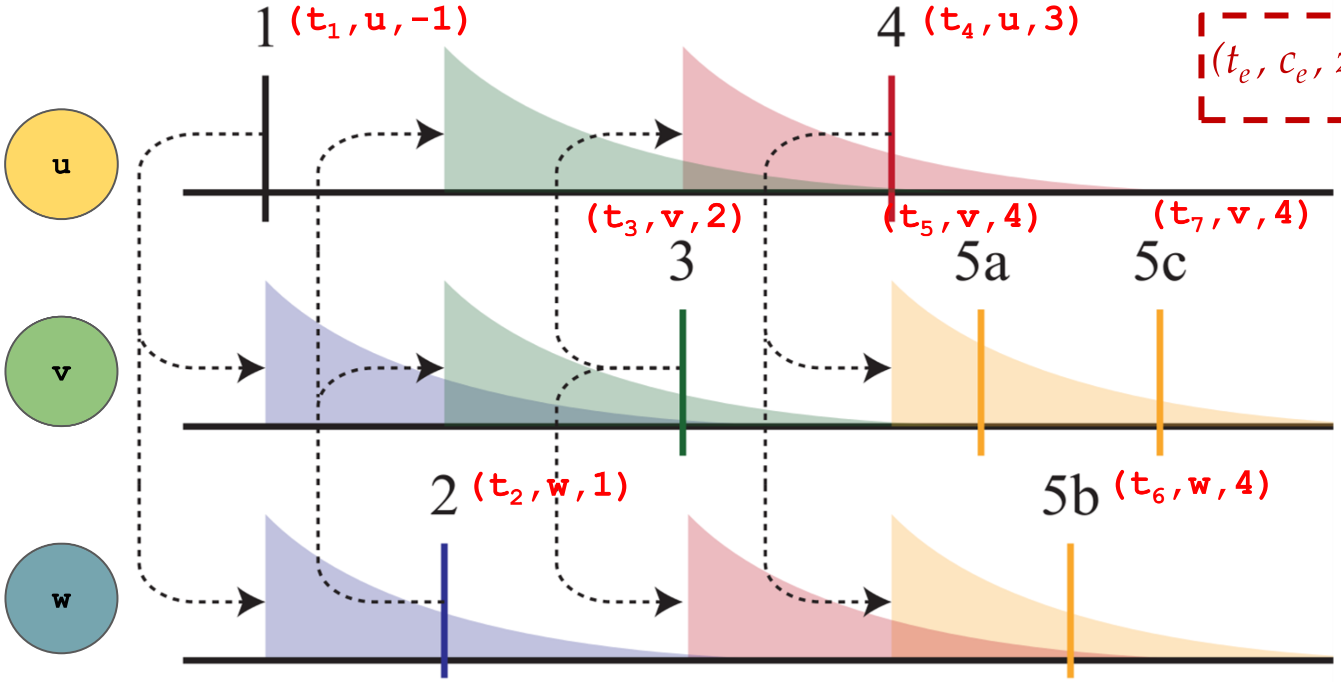
$$h_{c_n,v}(t - t_n) = \underbrace{W_{c_n v}}_{\text{User-User Influence}} \overbrace{f(\Delta t)}^{\text{Time Kernel}}$$

Explain

Multivariate Hawkes Process (MHP): *(HMHP, HTM, NetHawkes)*

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

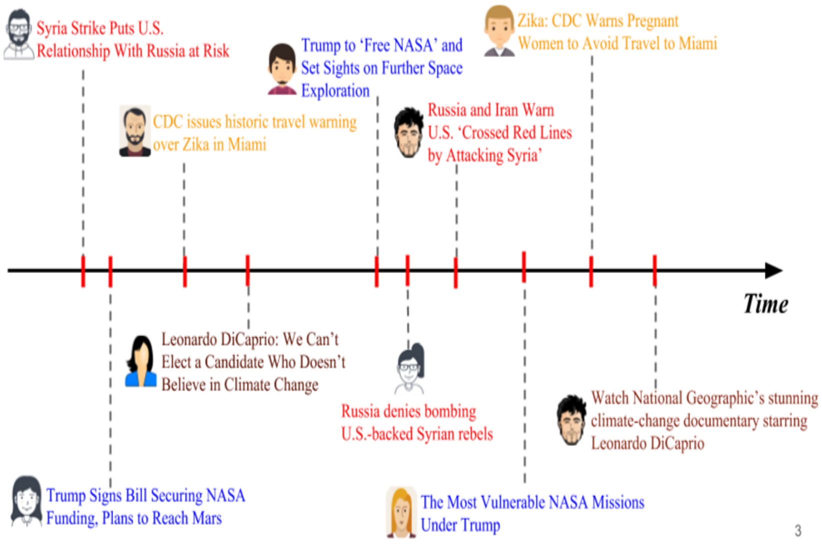
$$h_{c_n,v}(t - t_n) = W_{c_n v} f(\Delta t)$$



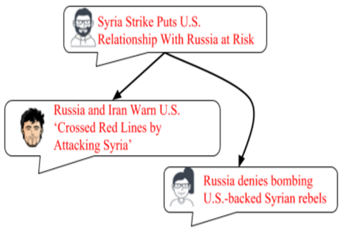
$(t_e, c_e, z_e) = (time, user, parent)$

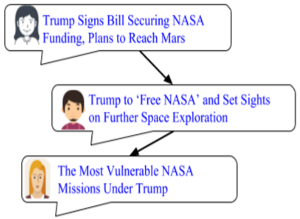
For each event, topics are sampled later independent of the time-stamps

What HTM [He. et al., 2015] does?

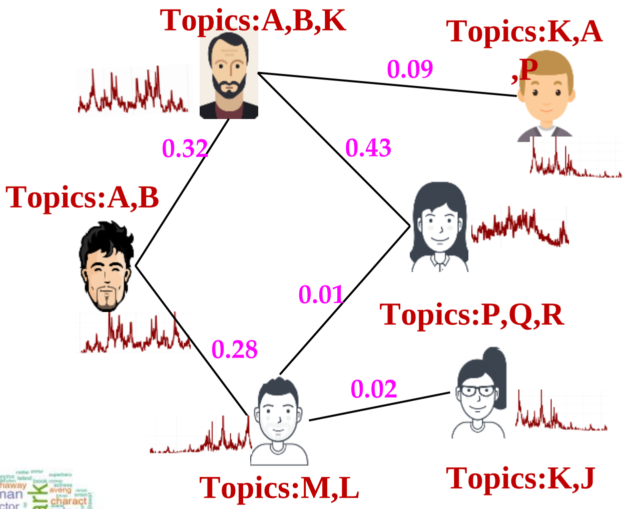


Set of tweets and time-stamps





Cascade Reconstruction



Network Reconstruction and Temporal Dynamics

Topic Model

Missing Topical Interactions in HTM *[He. et al., 2015]*

"If an event is triggered by another event, its document should be similar to the document of the triggering event. This suggests that the content of the user's post, influenced by the friend's previous post should have similar content to her friend's post." ***~HTM [He. et al., 2015]***

Parent-Child tweet pair

[#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** extracted by **HMHP

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

Missing Topical Interactions in HTM *[He. et al., 2015]*

Generation of Topic of child event in HTM

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

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

v/s

Generation of Topic of child event in HMHP

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

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

where, ζ is Topical Interaction Distribution

Parent-Child tweet pair

[#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.

How HMHP does this?

- *Coupling of Network MHP and the Markov Chain over topics.*
- *Coupled inference: Collapsed Gibbs sampling*

Generative Model

t_e= time, c_e=user, z_e= parent

1) Generate (t_e, c_e, z_e) for all events according Multivariate Hawkes Process.

Temporal Dynamics & Network Inference using MHP

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 $\boldsymbol{\phi}_v \sim Dir_K(\boldsymbol{\gamma})$

Topic-word, Topic-Topic, and User-Topic distributions resp.

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 $\eta_e \sim Discrete_K(\boldsymbol{\mathcal{T}}_{\eta_{z_e}})$

   Topic

   Topical Interaction Matrix

Cascade reconstruction and Topical Interactions coupling MHP and Topical MCs

   b) Sample document length $N_e \sim Poisson(\lambda)$
   c) For $w = 1 \ldots N_e$: draw word $x_{e,w} \sim Discrete_{\mathcal{W}}(\boldsymbol{\zeta}_{\eta_e})$

Topic Model

Inference : HMHP

Likelihood MTPP

$$\mathbb{P}(\mathcal{H}_T) := \left(\prod_{e_i \in \mathcal{H}_T} \overbrace{\lambda_{v_i}(t_i)}^{\text{Prob. of an action at } t_i} \underbrace{m^*(\eta_i)}_{\text{Prob. of mark } \eta_i} \right) \prod_{v \in V} \overbrace{\exp\left(\int_0^T \lambda_v(\tau) d\tau \right)}^{\text{Prob. of no actions at } t \in [0,T] \setminus \{t_i\}}$$

***Note**: The timestamps t_i and the marks(topics) η_i are modeled independently.*

Gibbs Sampling

- Suppose $p(x,y)$ is a p.d.f. that is difficult to sample from directly.
- But, can sample from the conditional distributions $p(x|y)$ and $p(y|x)$

Gibbs Sampling

- Set x and y to some initial values - say (x_0, y_0)
- For $i = 1$ to M(#iterations):
    - Sample $x_i \sim p(x|y_{i-1})$
    - Sample $y_i \sim p(y|x_i)$

Parameters to infer: HMHP

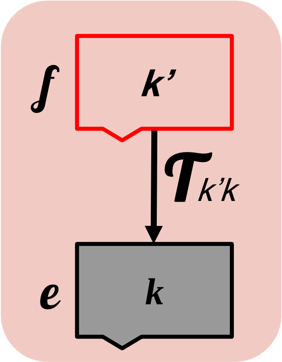
- Topic for each event
- Parent for each event
- User-User influence (for all pairs of users)
- ~~*Topic-Topic interaction?*~~ *(Integrates out due to Collapsed Gibbs Sampling)*

Cascade Inference (Parent Assignment)

Time Stamps

$Pr(par(e) = f \mid Topics, W, \{ti\}) \propto$

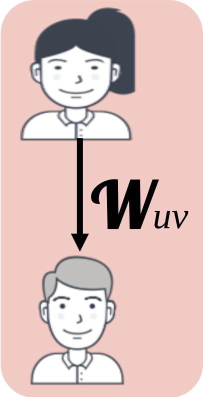
User-User Influence



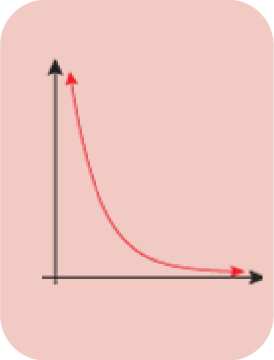
Topical interaction between parent-child events

*

Interaction between the users of the par-child events



*



Exponential decay (Time kernel)

Topic Inference

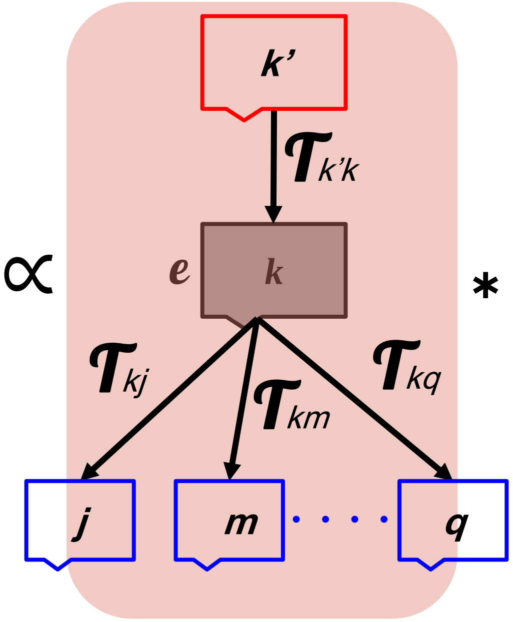
content of the event e

$$Pr(topic(e) = k \mid \{z_1, z_2, \ldots\}, \{w_1, w_2, \ldots\}, Topics \setminus \{topic(e)\}) \propto$$

parent assignment of all events

topic assignment of all other events

Topical interaction at the parent child structure



*

The Most Vulnerable NASA Missions Under Trump

Content of event e

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

Results : HMHP

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.*

HMHP performs better on Semi-Synthetic dataset

HMHP Anecdotal Results : *Real Dataset*

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

Generative Model : DNHP

What is missing in HMHP? *(as well as in HTM, NHWKS)*

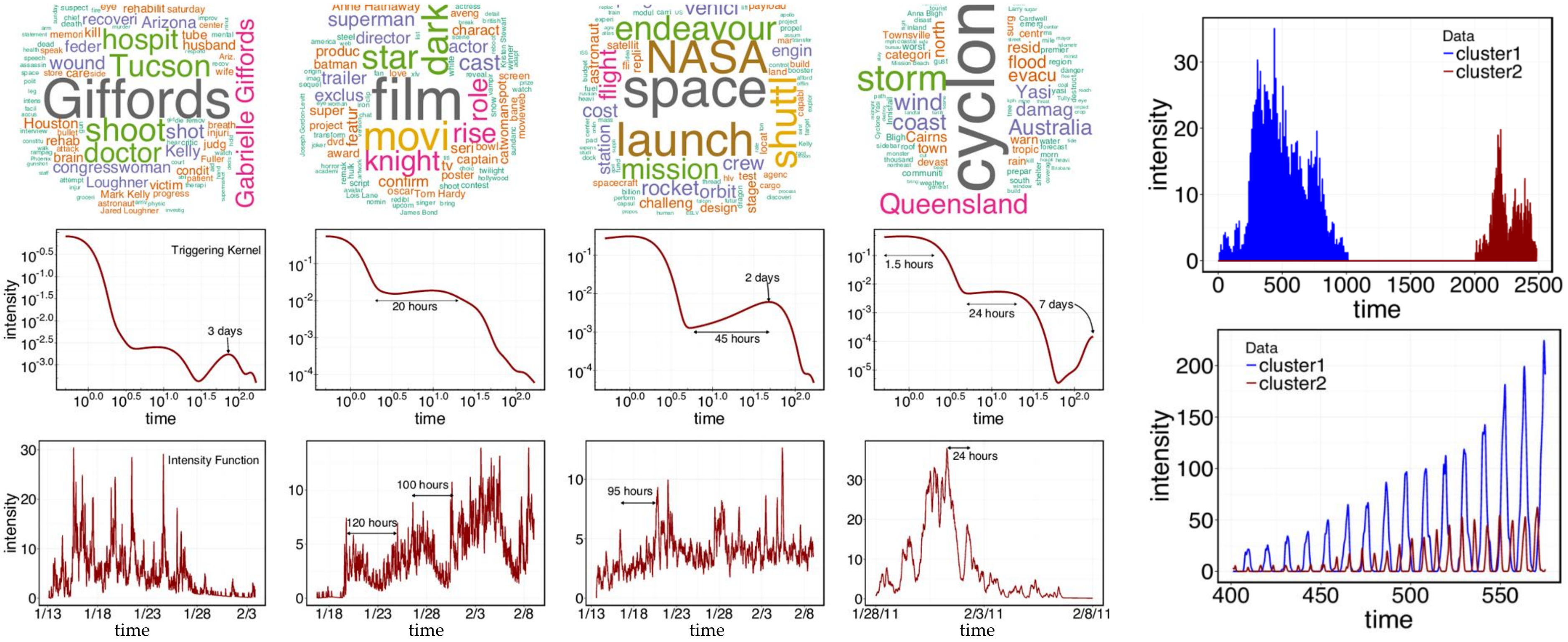
$$\lambda_v(t) = \mu_v(t) + \sum_{n=1}^{|\mathcal{H}_{t-}|} h_{c_n,v}(t - t_n)$$

$$h_{c_n,v}(t - t_n) = W_{c_n v} f(\Delta t)$$

If user v likes user c_n, it would try generating a time-stamp

***Note** that this **does not** take the into **account** the **topic** of the **event** generated **by c_n**.*

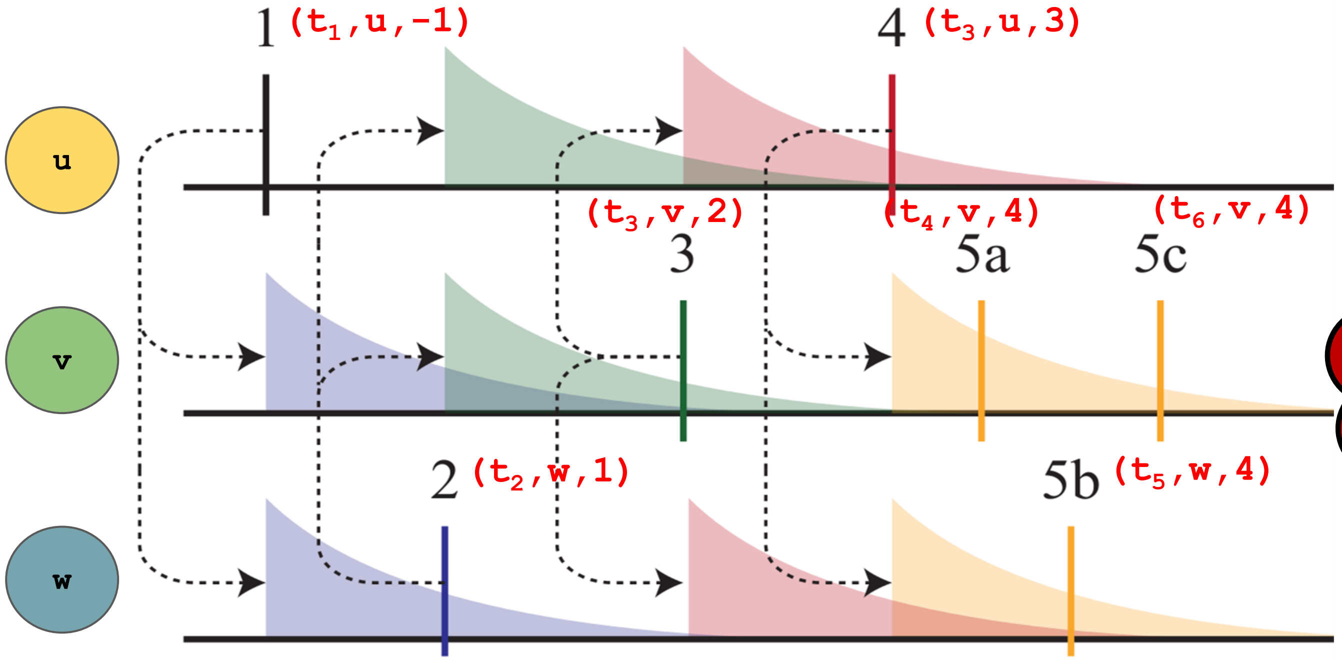
Time-Topic relation evidence: *[DirHawkes Du et al. ‘15]*



Multivariate Hawkes Process (MHP): *(HMHP, HTM, NetHawkes)*

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

$$h_{c_n,v}(t - t_n) = W_{c_n v} f(\Delta t)$$



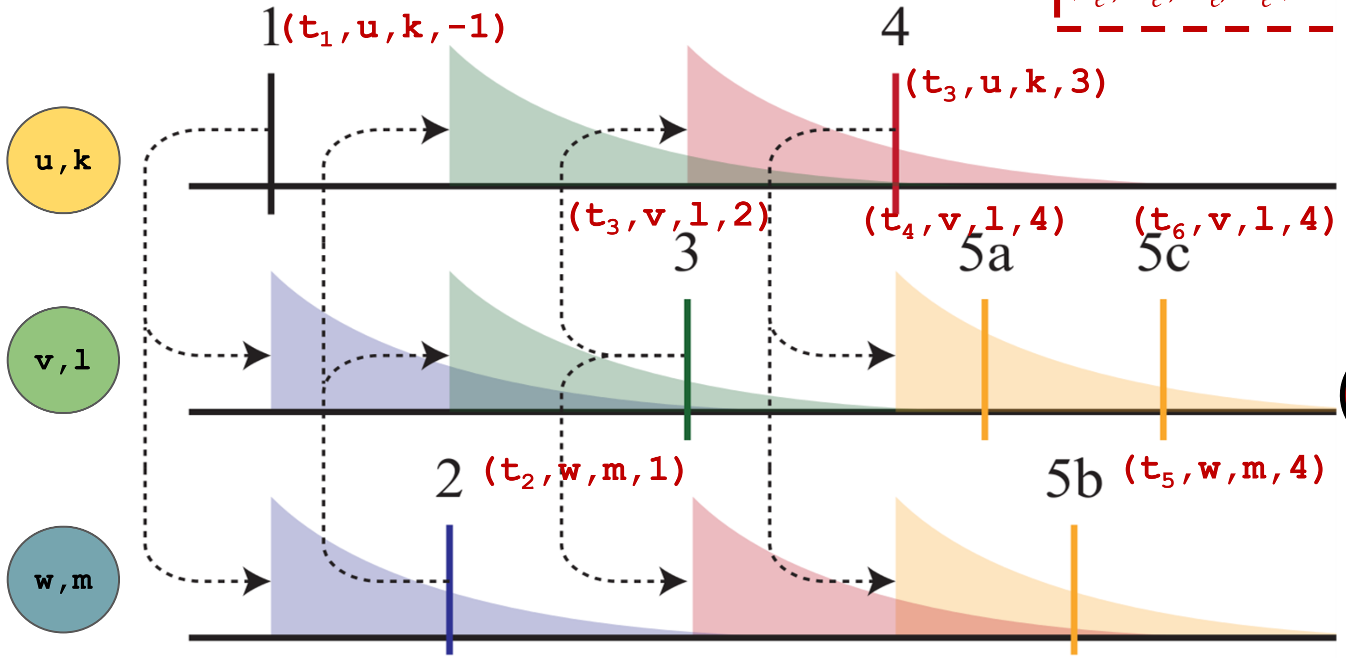
For each event, topics are sampled later independent of the time-stamps

Marked Multivariate Hawkes Process: *(DNHP)*

$$\lambda_{\mathbb{v}}(t) = \mu_{\mathbb{v}}(t) + \sum_{n=1}^{|\mathcal{H}_{t-}|} h_{\mathbb{c}_n,\mathbb{v}}(t - t_n)$$

$$h_{c_n,v}(t - t_n) = W_{uv} T_{kk'} f(\Delta t)$$

$(t_e, c_e, k_e, z_e) = (time, user, topic, parent)$



For each event, topic comes along with the event generation time.

What we add to HMHP → DNHP?

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

$$h_{c_n,v}(t - t_n) = W_{c_n v} f(\Delta t)$$

If user v likes user c_n, it would try generating a time-stamp

***Note** that this **does not** take the into **account** the **topic** of the **event** generated **by c_n**.*

HMHP

DNHP

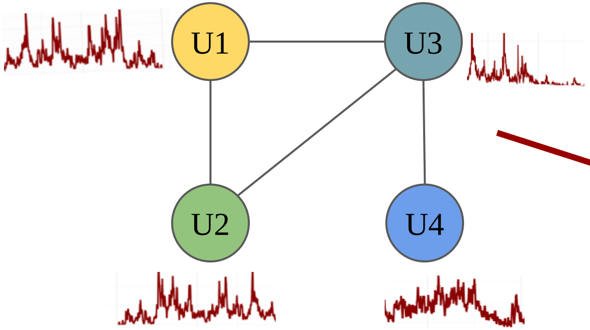
$$\lambda_{\mathbb{v}}(t) = \mu_{\mathbb{v}}(t) + \sum_{n=1}^{|\mathcal{H}_{t-}|} h_{\mathbb{c}_n,\mathbb{v}}(t - t_n)$$

$$h_{c_n,v}(t - t_n) = W_{uv} T_{kk'} f(\Delta t)$$

If user v likes user c_n, and also the topic of event by c_n, then it would try generating a time-stamp

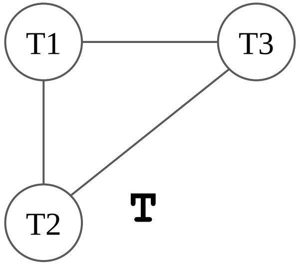
HMHP v/s DNHP

HMHP



A

Hawkes process for each user node

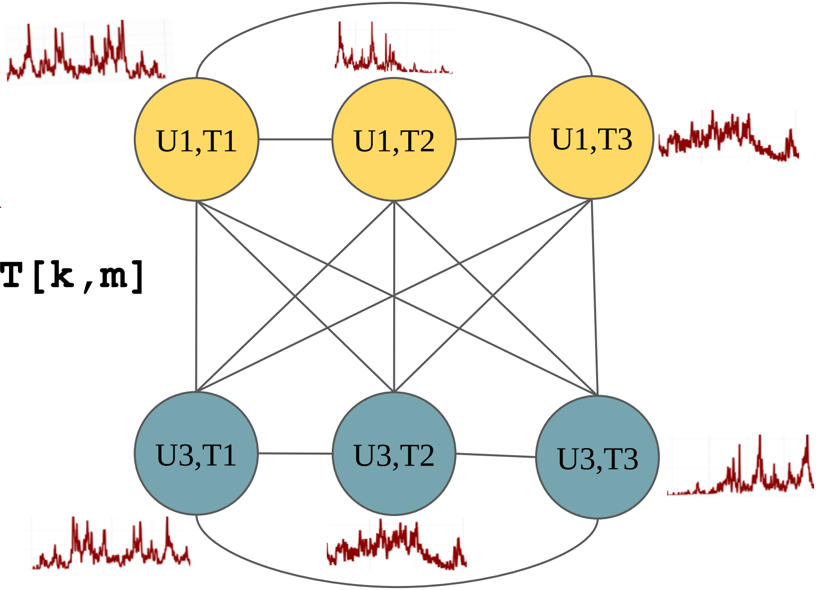


T

`A[u,v]*T[k,m]`

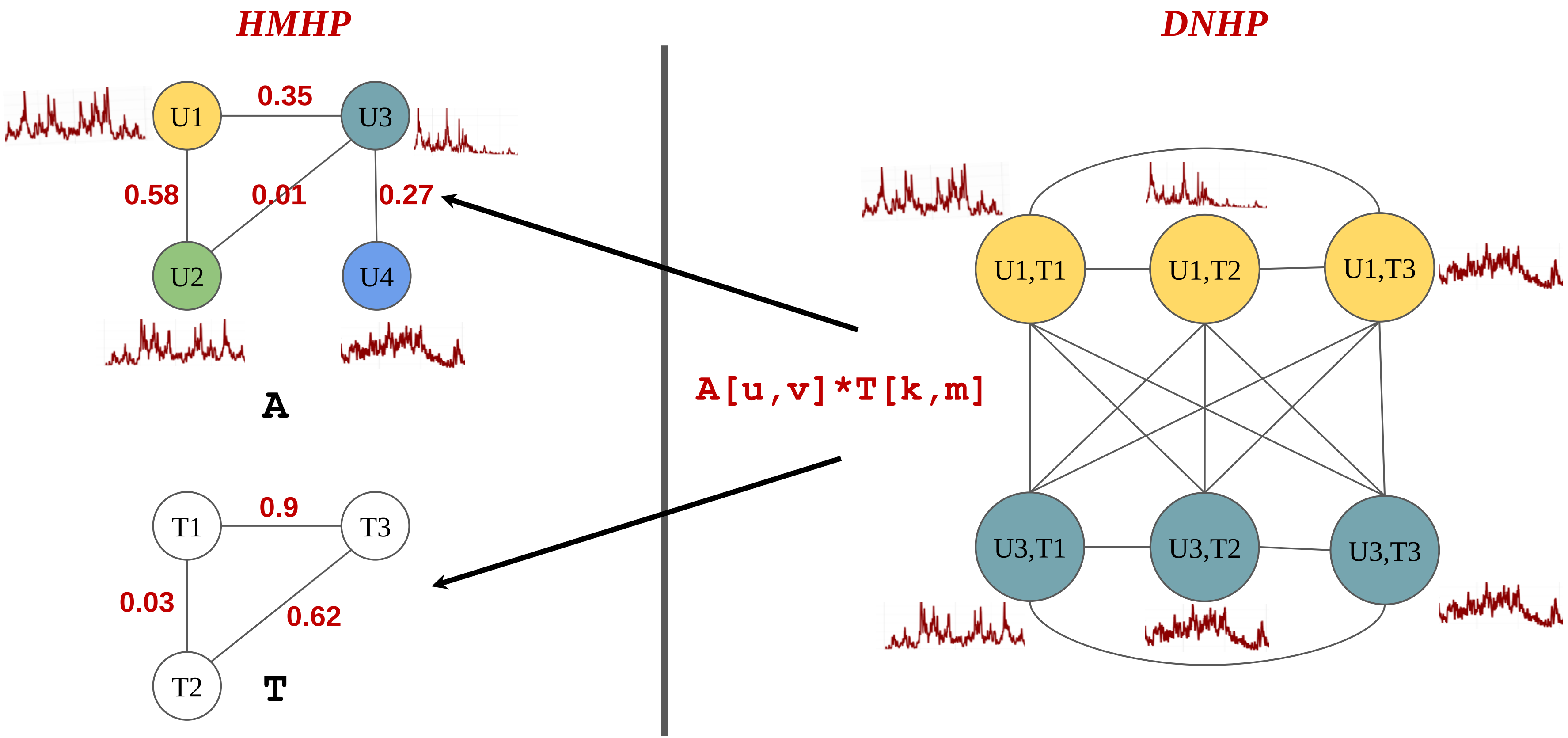
DNHP

Hawkes process for each user-topic pair node



Topics are still latent!

HMHP v/s DNHP



HMHP
DNHP
U1
U2
U3
U4
0.35
0.58
0.01
0.27
A
T1
T2
T3
0.9
0.03
0.62
T
A[u,v]*T[k,m]
U1,T1
U1,T2
U1,T3
U3,T1
U3,T2
U3,T3

DNHP Likelihood Form

$$\mathbb{P}\left(\mathcal{H}_T\right) := \left(\prod_{e_i \in \mathcal{H}_T} \overbrace{\lambda_{\mathbb{v}_i}(t_i)}^{\text{Prob. of an action at } t_i \text{ with mark } \eta_i} \right) \prod_{\mathbb{v} \in \mathcal{V}} \overbrace{\exp\left(\int_0^T \lambda_{\mathbb{v}_i}(\tau) d\tau \right)}^{\text{Prob. of no actions at } t \in [0,T] \setminus \{t_i\}}$$

$$\mathbb{v}_i = (v_i, \eta_i)$$

Recall:

$$\mathbb{P}\left(\mathcal{H}_T\right) := \left(\prod_{e_i \in \mathcal{H}_T} \overbrace{\lambda_{v_i}(t_i)}^{\text{Prob. of an action at } t_i} \underbrace{m^*(\eta_i)}_{\text{Prob. of mark } \eta_i} \right) \prod_{v \in V} \overbrace{\exp\left(\int_0^T \lambda_v(\tau) d\tau \right)}^{\text{Prob. of no actions at } t \in [0,T] \setminus \{t_i\}}$$

HMHP (HTM,NetHawkes) v/s DNHP

HMHP (HTM, NetHawkes)

DNHP

Intensity Function

(each user node)

v

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

(a user-topic pair)

$\mathbb{v} = v,k$

$$\lambda_{\mathbb{v}}(t) = \mu_{\mathbb{v}}(t) + \sum_{n=1}^{|\mathcal{H}_{t-}|} h_{\mathbb{c}_n,\mathbb{v}}(t - t_n)$$

Base Rate

v

$$\mu_v(t) = \mu_v(t)$$

$\mathbb{v} = v,k$

$$\mu_{\mathbb{v}}(t) = \mu_v(t)\mu_k(t)$$

Impulse Response

$u \rightarrow v$

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

$u,k \rightarrow v,k$

$$h_{\mathbb{u},\mathbb{v}}(\Delta t) = W_{u,v} \mathcal{T}_{k,k'} f(\Delta t)$$

HMHP (HTM,NetHawkes) v/s DNHP

HMHP (HTM, NetHawkes)

DNHP

Influence Inference

$$W_{u,v} \sim Gamma\left(N_{u,v} + \alpha_1, N_u + \beta_1\right)$$

Note: *Topic-Topic interaction is integrated out in HMHP because of conjugacy*

Coupled/interacting parameters

$$W_{u,v} \sim Gamma\left(N_{u,v} + \alpha_1, \sum_k \left(N_{u,k} \sum_{k'} \mathcal{T}_{k,k'}\right) + \beta_1\right)$$

$$\mathcal{T}_{k,k'} \sim Gamma\left(N_{k,k'} + \alpha_1, \sum_u \left(N_{u,k} \sum_{v} W_{u,v}\right) + \beta_1\right)$$

Base Rate Inference

$$\mu_v = \frac{N_v^{(spon)}}{T}$$

Note: *There is no base rate associated with the topics*

$$\mu_v = \frac{N_v^{(spon)}}{T \sum_{k \in K} \mu_k} \qquad \mu_k = \frac{N_k^{(spon)}}{T \sum_{v \in V} \mu_v}$$

Results : DNHP

Datasets

Twitter (Real Data):

- ***Tweets*** *from* ***151 US Congress Members*** *comrpising of 360K tweets* **--** *gathered using Twitter API in July 2018.*

Semi-Synthetic:

- *Retain the underlying set of nodes and the follower graph from Twitter Data.*
- *Estimate the parameters required for our model from the data.*
- *Generate* ***5 different samples*** *of* ***360K events*** *using* ***DNHP*** *model.*

DNHP performs better on Semi-Synthetic dataset

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*
- **HMHP**
- **DNHP**

DNHP Generalization Performance: *Real Dataset*

AVG. LL (TIME + CONTENT)				
#Topics = 25				
TRAIN (TEST)	DNHP	HMHP	HWK+DIAG	HWKxLDA
114K (70K)	-80.51	-95.71	-100.21	-96.53
177K (100K)	-78.09	-86.66	-91.06	-87.34
240K (130K)	-76.57	-80.14	-83.96	-80.18
#Topics = 100				
114K (70K)	-80.51	-95.71	-102.16	-97.46
177K (100K)	-78.09	-86.66	-92.99	-88.16
240K (130K)	-76.57	-80.14	-85.64	-80.88

Higher the better

DNHP Generalization Performance: *Real Dataset*

AVG. LL (TIME)				
	#TOPICS = 25		#TOPICS = 100	
TRAIN (TEST)	DNHP	NHWKS	DNHP	NHWKS
114K (70K)	-8.74	-24.46	-8.04	-24.46
177K (100K)	-7.45	-16.32	-7.09	-16.32
240K (130K)	-6.56	-10.27	-6.37	-10.27

Higher the better

67.13%
56.55%
37.97%

The **gap** between DNHP and HMHP is **larger** for **smaller dataset size (the training size)**.

Parameters are **learned efficiently** using **flow of evidence between parameters**.

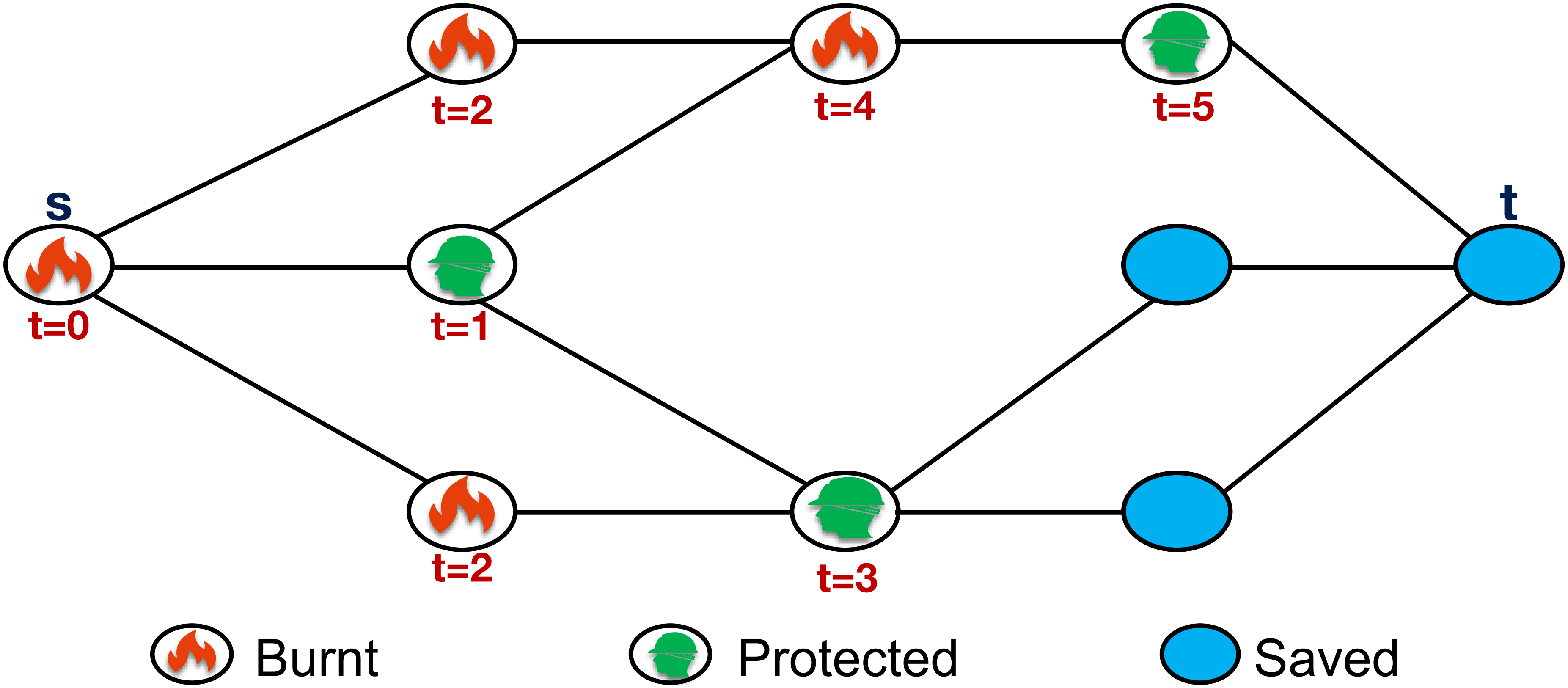
Conclusion

- *HMHP & DNHP, account for* ***topical interactions, user-user influence, user-topic patterns****.*
- *In DNHP the* ***rate of generation of events*** *is* ***dependent*** *not only on the* ***users*** *but also on the* ***topic or mark*** *associated with the event.*
- *In DNHP,* ***topical interactions*** *&* ***user-user influence*** *are* ***coupled****, and joint estimation of these parameters enables* ***flow of evidence*** *across the parameters.*
- *In both HMHP & DNHP, incorporating* ***topical interactions*** *and the* ***collective inference*** *of parameters leads to* ***more accurate estimation latent parameters****, also,* ***fits the real Twitter conversations better*** *(in terms of test likelihood) as compared to other state-of-the-art models.*

Part-II: Algorithmic Perspective

Saving a Critical Set with Firefighters

The Firefighting Problem



What is there to do?

- **Maximizing** the **spread of influence** *[Kempe et al. '03]*
- **Minimizing** the **fraction of infected population**, **minimizing** the **time** of **detection** of infection *[Leskovec et al. '07, Ceren et al. '11]*

- **Max**imising the number of **saved vertices** *[Cai, Verbin, and Yang, '08]*
- **Min**imising the number of **burned (infected) vertices** *[Cai, Verbin, and Yang, '08, Finbow, Hartnell, et. al., '09]*
- **Min**imising the number of **rounds**, **min**imising the number of **firefighters per round** *[Anshelevich, Chakrabarty, et. al., '09]*
- **Saving** a **specific set of vertices** *[King, MacGillivray, '09]*

❖ ***Saving a Critical Set with Firefighters is FPT [Choudhari et al. '17]***

Saving a Critical Set

Saving a Critical Set (SACS)

Input: *An undirected n-vertex graph G, a vertex s, a subset $C \subseteq V(G) \setminus \{s\}$, and an integer k*

Question: *Is there a valid k-step strategy that saves C when a fire breaks out at s?*

Basic Definitions

Fixed Parameter Tractability

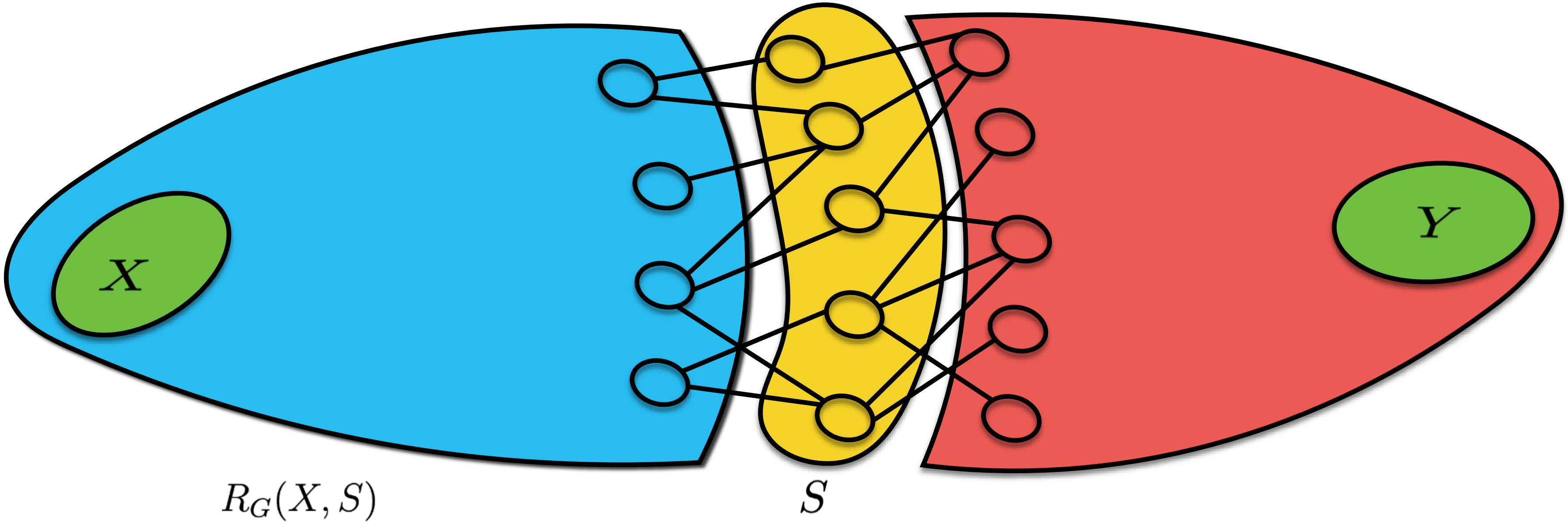
Definition (Parameterized Problem):

*A **parameterization** of a decision problem is a function that assigns an **integer parameter k** to each **input instance I**.*

Definition (FPT):

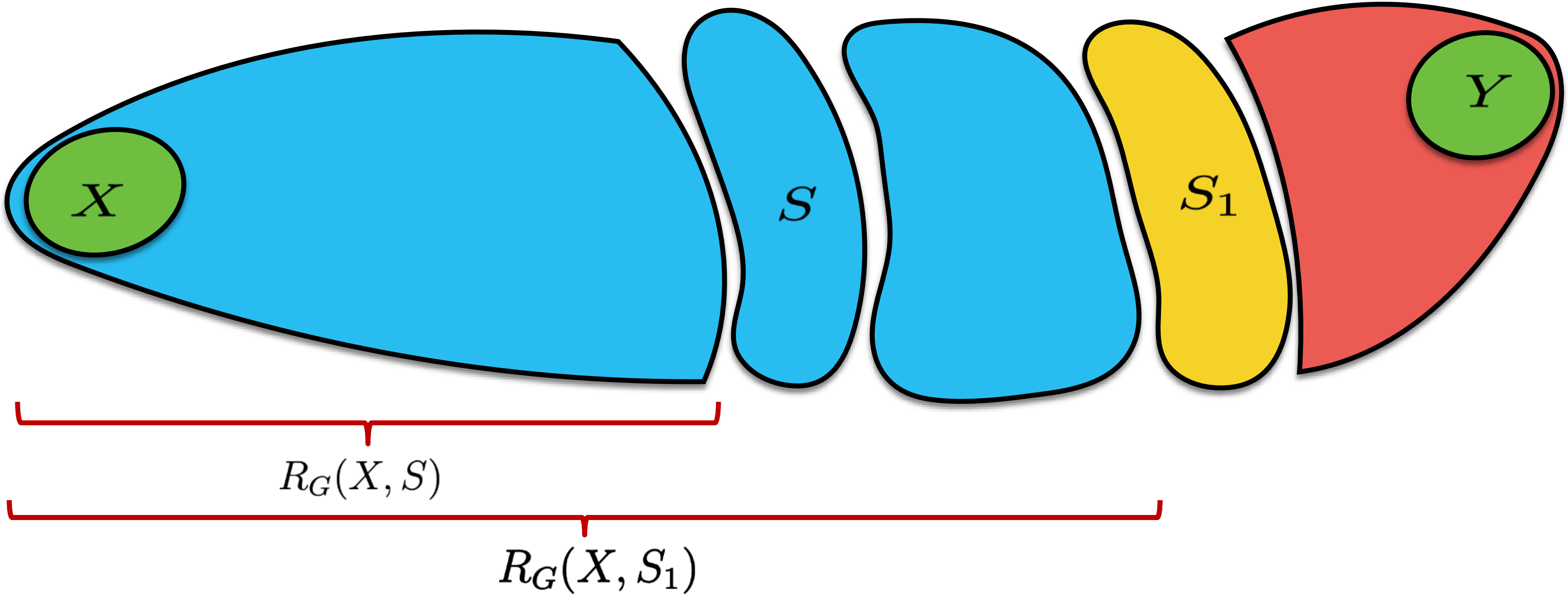
*A parameterized problem is **fixed-parameter tractable (FPT)** if there is an **$f(k)n^c$** time algorithm for some **constant c**.*

Separator



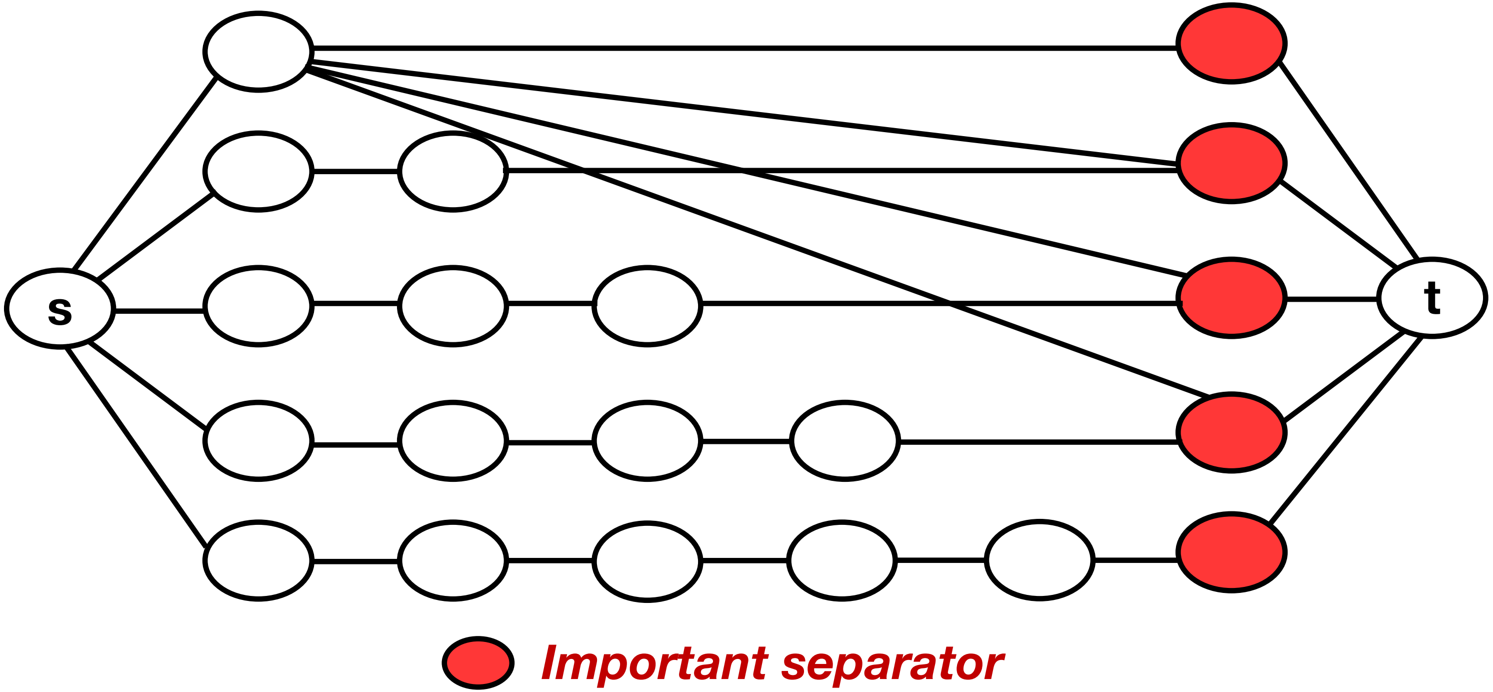
Reachable Set of X w.r.t. S

Dominating Separator



- $|S_1| \leq |S|$
- $R_G(X,S) \subseteq R_G(X,S_1)$

Important Separator



*Important separators are those which are **not dominated** by any other separator*

Firefighting on Trees

Saving a Critical Set (SACS) on Trees

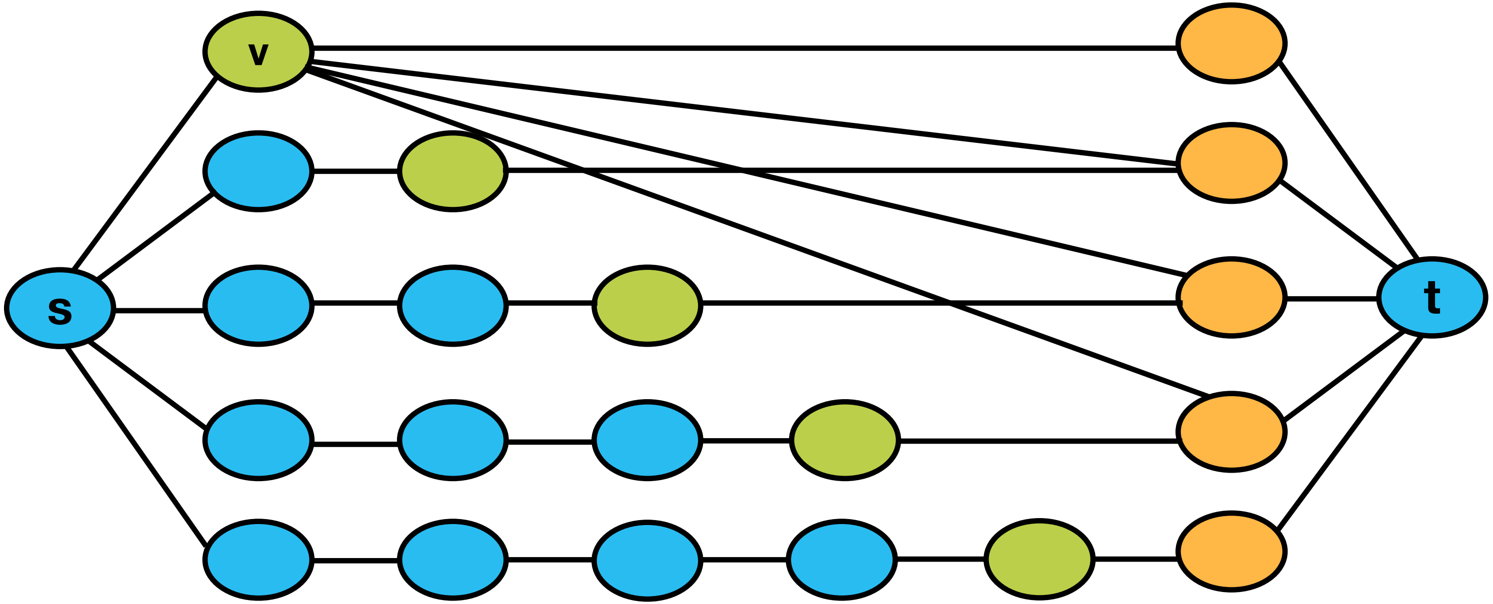
Theorem: (Marx, '11)

*For **trees**, there are **at most 4^k important separators** of size at most **k**.*

SACS on trees takes time $O^*(4^k)$

Firefighting on Graphs

Firefighting on Graphs with Important Separators



Firefighting Solution

Important Separator

Important separators do not suffice !!!

Saving a Critical Set: Para-NPC

Saving A Critical Set ***(SACS)*** with ***critical set of size 1*** is a ***YES-instance***

if and only if

k-CLIQUE is an ***YES-instance***

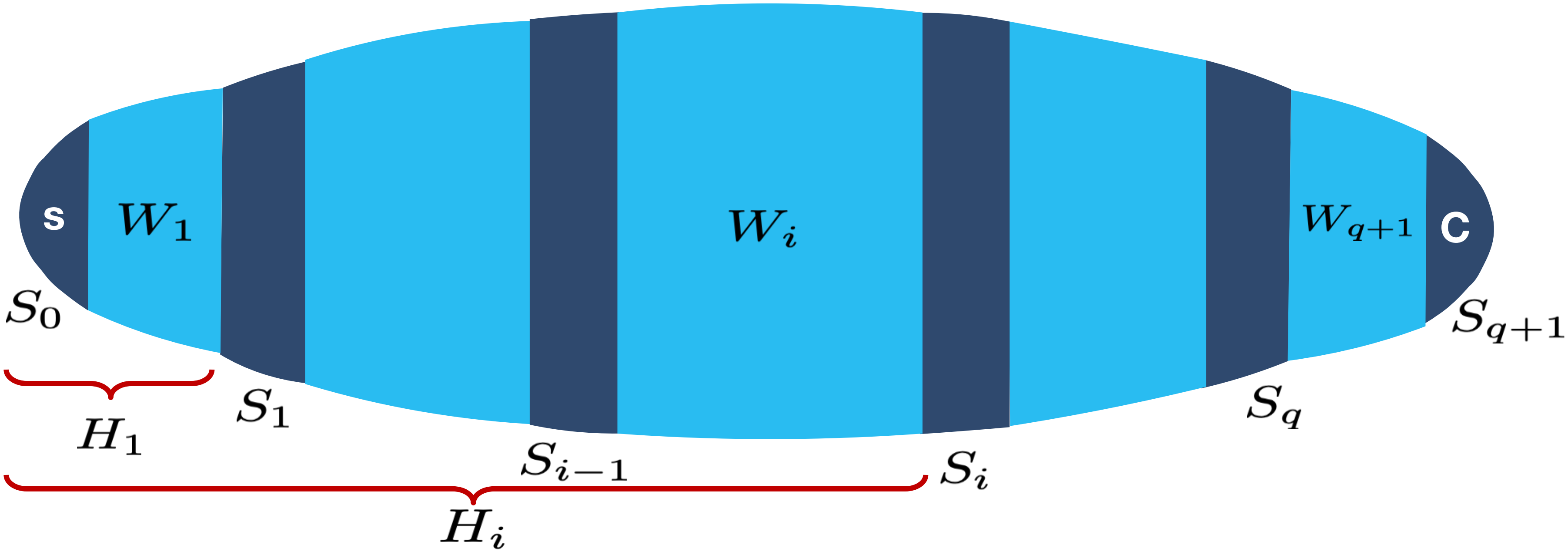
Saving a Critical Set: Para-NPC

SACS with critical set of size 1 *has a successful strategy with* ***$(k + m - {}^{k}C_2)$ firefighters*** *in this new graph* ***G′***
if and only if
G *has a* ***clique of size k***

Proof

FPT Solution on Graphs

Tight Separator Sequence (Formal Definition)



*There is an algorithm that runs in time **O(kmn²)** that either correctly concludes that there is no **X-Y separator** of size at most **k** or outputs the required sequence.*

*[*Lokshtanov et al. '16*]*

Overview of the FPT Algorithm

1. *Compute a sequence of separators (<u>Tight Separator Sequence</u>)* ***(bounded in poly n)***

2. ***Consider** a **behavior (labeling) of the firefighting solution** on the nodes **in** these **separators***

3. *Consider **two consecutive separators and the region between** them along with the **labelled firefighting solution**. Call it as **border problem***

4. ***Repeat** for all the **consecutive border problems (bounded in k)**. If **all** the **border problems** return **YES**, then **Algo** return **YES**. **Else,***

5. *Go for the **new behavior** (labeling) i.e. Step-2 – **(#behaviors are bounded in poly k)**.*

Case 1: $q > k$

Let

$$S = \bigcup_{i=1}^{q} S_i \qquad\qquad W = \bigcup_{i=1}^{q+1} W_i$$

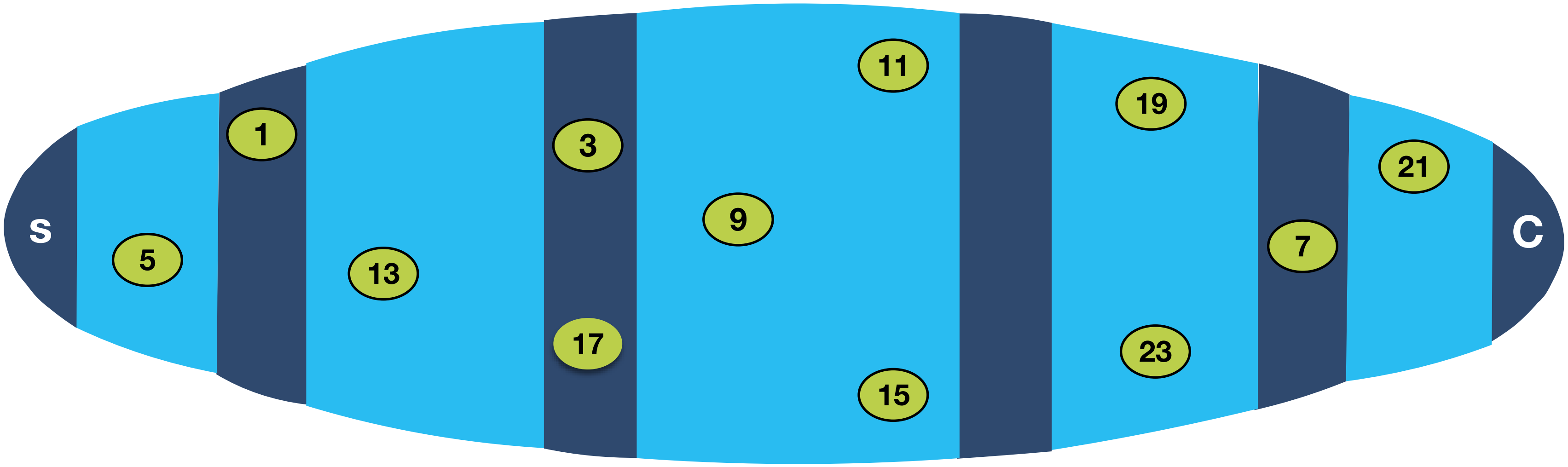
Claim: *If G admits a tight (s,C)-separator sequence of order q in G \ Y where q > k, then there exists a k-step firefighting strategy.*

$$\boxed{Place\ the\ firefighters\ on\ the\ separator = S_q}$$

Case 2: $q \leq k$

Guess the partition of the timestamps P for a firefighting strategy

For e.g., $P = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23\}$



$\{5\}$ $\{1\}$ $\{13\}$ $\{3, 17\}$ $\{9, 11, 15\}$ $\{\ \}$ $\{19, 23\}$ $\{7\}$ $\{21\}$

Partition over time-stamps

Let,

- $A_1, A_2, \ldots, A_q$ $denote\ timestamps\ for\ nodes\ inside\ S\ and$
- $B_1, B_2, \ldots, B_{q+1}$ $denote\ timestamps\ for\ nodes\ inside\ W$

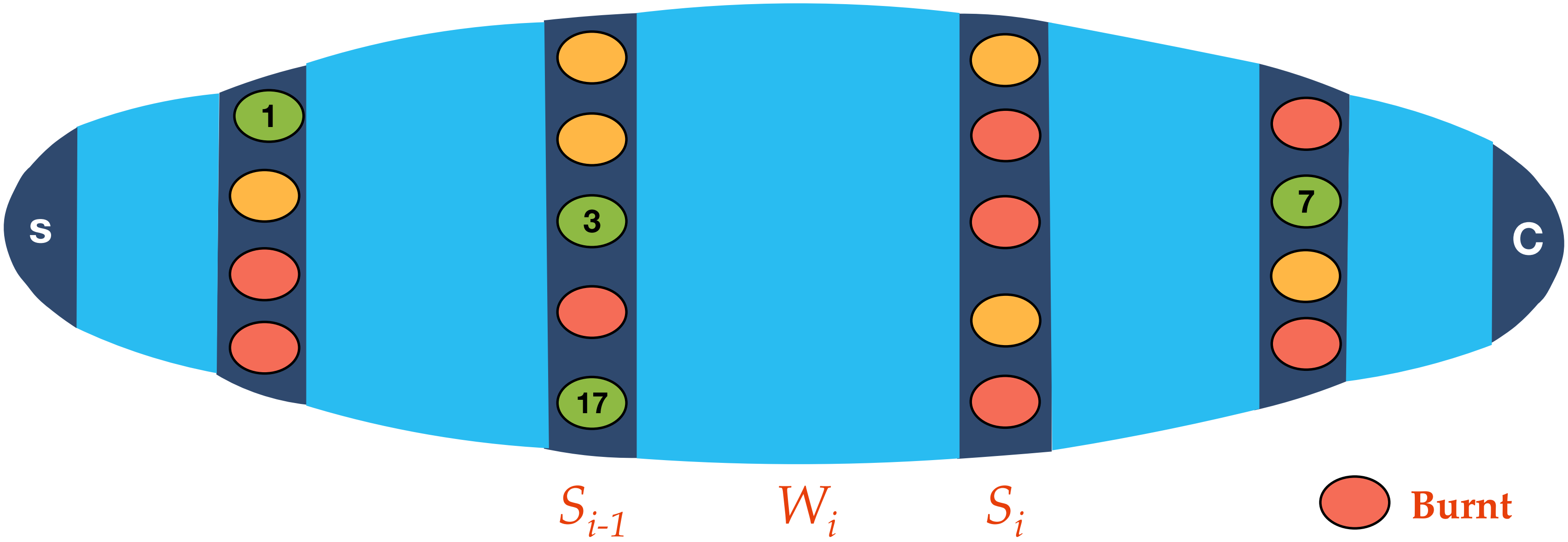
$$P = \bigcup_{i=1}^{q} A_i \cup \bigcup_{i=1}^{q+1} B_i$$

$$|P| = p$$

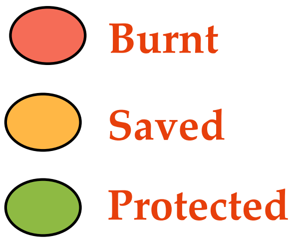
$$The\ number\ of\ possible\ partitions = (2q+1)^p \leq (2k+1)^k$$

Possible Labeling

Guess the behaviour of the strategy restricted to $S = \cup_{i=1\ldots q} S_i$



$$G_i = G[S_{(i-1)} \cup W_i \cup S_i]$$



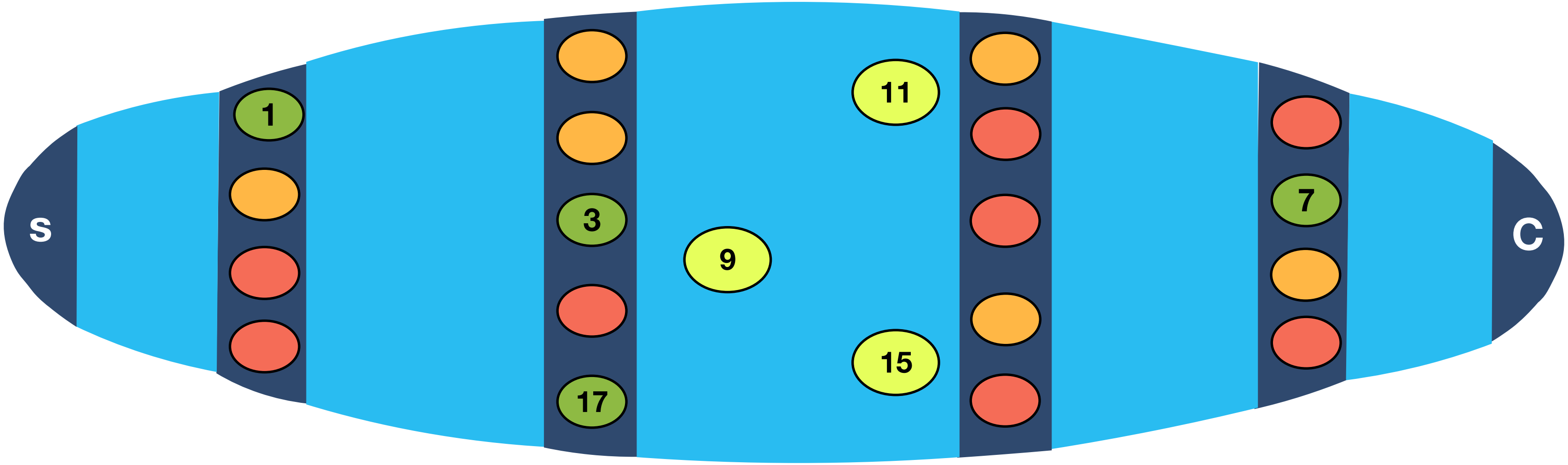
Possible Labelings

$$\mathfrak{L} = (\{\mathfrak{f}\} \times X) \cup (\{\mathfrak{b}\} \times [2k]_E) \cup \{\mathfrak{p}\}$$

$$\mathfrak{L}_{\mathfrak{h}}(v) = \begin{cases} (\mathfrak{f}, t) & \text{if } \mathfrak{h}(t) = v, \\ (\mathfrak{b}, t) & \text{if } t \text{ is the earliest timestep at which } v \text{ burns}, \\ \mathfrak{p} & \text{if } v \text{ is not reachable from } s \text{ in } G \setminus (\{\mathfrak{h}(i) \mid i \in [2k]_O\}) \end{cases}$$

$\boxed{\textit{The number of possible partitions} = (k+k+1)^{k*k} \leq (3k)^{k^2} \leq k^{O(k^2)}}$

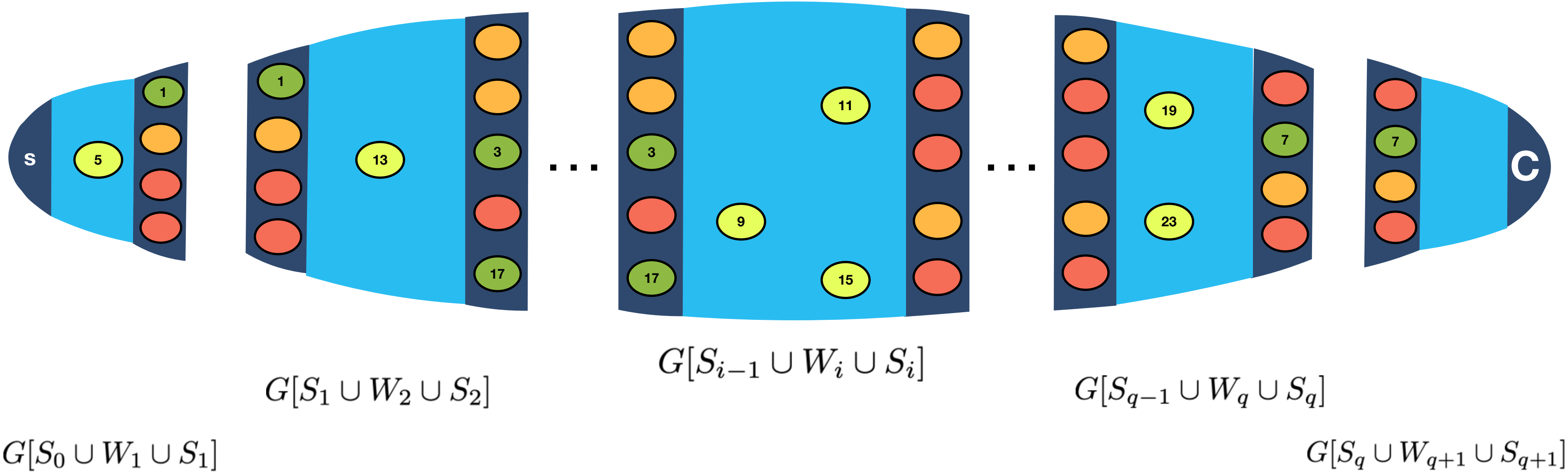
The Border Problem : Solve Recursively



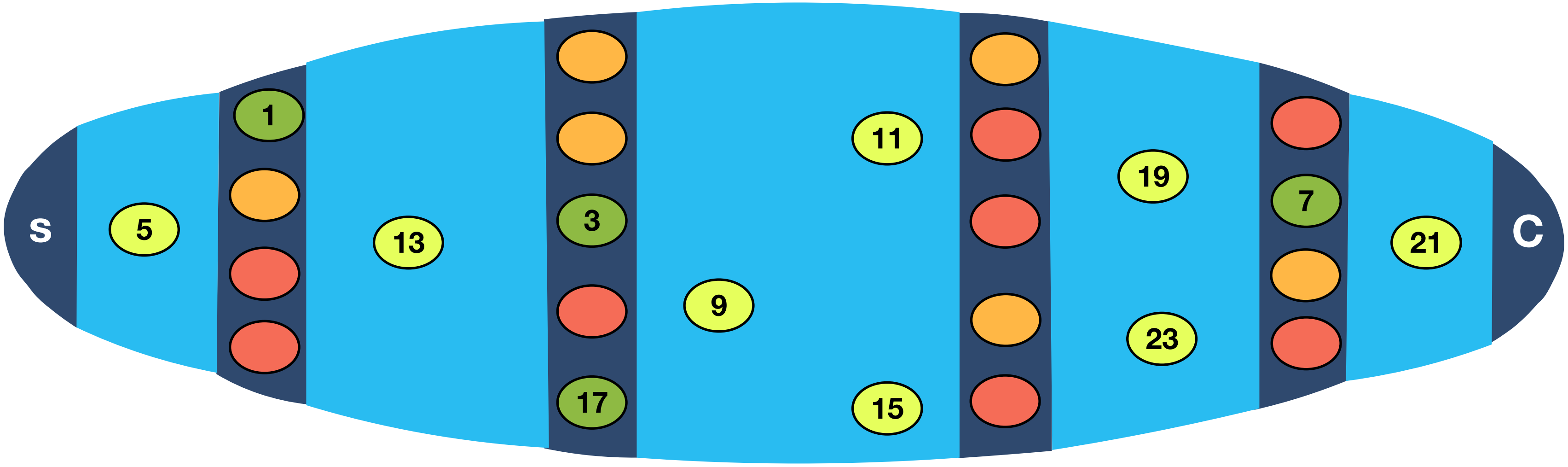
S_{i-1} W_i S_i

$$G_i = G\,[S_{(i-1)} \cup W_i \cup S_i]$$

The Border Problem : Solve Recursively



Combining the Solution



Patch all the Border Problems together

Algorithm

Algorithm 1: Solve-SACS-R($\mathcal{I}$)

Input: An instance $(G, s, C, k, g, P, Q, Y, \gamma)$, $p := |P|$

Result: YES if $\mathcal{I}$ is a YES-instance of SACS-R, and NO otherwise.

if *$p = 0$ and s and C are in different components of $G \setminus Y$* **then return** YES;

else return NO;

if *$p > 0$ and s and C are in different components of $G \setminus Y$* **then return** YES;

if *there is no $s - C$ separator of size at most p* **then return** NO;

Compute a tight $s - C$ separator sequence $\mathcal{S}$ of order p.

if *the number of separators in $\mathcal{S}$ is greater than k* **then return** YES;

else

    for *a non-trivial partition $\mathcal{T}_1(P), \mathcal{T}_2(P)$ of P into $2q+1$ parts* **do**

        for *a labeling $\mathfrak{T}$ compatible with $\mathcal{T}_1(P)$* **do**

            if $\bigwedge_{i=1}^{q+1}(\text{Solve-SACS-R}(\mathcal{I}\langle i, \mathcal{T}_1(P), \mathcal{T}_2(P), \mathfrak{T}_i\rangle))$ **then return** YES;

    return NO

Running Time

$$T(n, m, k, p) \leq O(n^2 mp) + (p + k + 1)^{kp} \sum_{i=1}^{q+1} T(n_i, m_i, k, p_i)$$

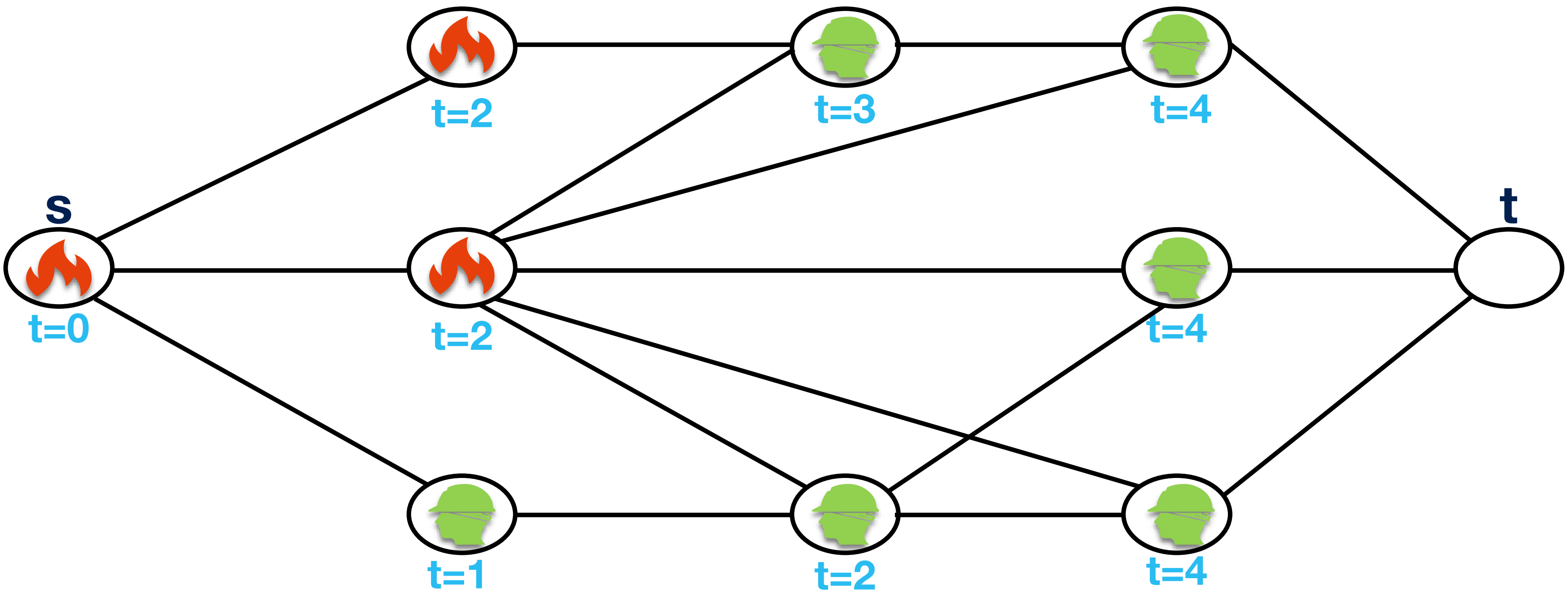
Recall that:

- *each* $p_i \leqslant k$,
- *the depth of recursion is bounded by p, and*
- *at each level, the work done is proportional to* $k^{O(kp)}\, n^2 m$

SACS** is **FPT** and has an algorithm with running time **f(k) O(n²m)
*where, **f(k) = k^O(poly k)**.*

The Spreading Model

Firefighting: The Spreading Model



Spreading Vaccination Model

*In the spreading model, **SACS** is hard as **k-DOMINATING SET***

Kernels on Trees

Kernelization

*A **kernelization** algorithm, or simply a **kernel**, for a **parameterized problem Q** is an **algorithm A** that, given an instance **(I,k) of Q**, works in polynomial time and returns an **equivalent instance** **(I′,k′) of Q**. Moreover, we require that* $k' \leq k$.

Firefighting on Trees: No poly Kernels

SACS when restricted to trees does not admit a polynomial kernel.

*The **unparameterized version of SACS** restricted to trees **cross composes** to **SACS** restricted to trees when **parameterized by the number of firefighters***

Conclusion & Future Work

Conclusion

- *HMHP & DNHP, account for **topical interactions, user-user influence, user-topic patterns**.*
- *In DNHP the **rate of generation of events** is **dependent** not only on the **users** but also on the **topic or mark** associated with the event.*
- *In DNHP, **topical interactions** & **user-user influence** are **coupled**, and joint estimation of these parameters enables **flow of evidence** across the parameters.*
- *In both HMHP & DNHP, incorporating **topical interactions** and the **collective inference** of parameters leads to **more accurate estimation latent parameters**, also, **fits the real Twitter conversations better** (in terms of test likelihood) as compared to other state-of-the-art models.*

- ***SACS** is **FPT** when **parameterized** by **number of firefighters**.*
- ***No polynomial** sized **kernel** for **trees**.*
- *In contrast to general Firefighting model, the **spreading model** is **W[2]-Hard**.*

Open Questions

- Sample complexity for Single Topic Model *[Arora et al. 2012, Bhattacharya Kannan, 2020]*
- Bayesian Non-Parametric
- Scalable inference (and log-likelihood calculation) for cascade-based models. (*Can this be framed as an FPT problem?)*
- Sketches to maintain high dimensional matrices *[Tassarotti et al. 2019]*
- Priors for incorporating correlation among parameters

Open Questions

- Probabilistic spread of fire
- Vaccination Strategies [Grauer et al., 2020]
- Maintaining separators in a streaming and/or dynamic settings
- Firefighter over insertion stream of edges and/or in a dynamic streams

Acknowledgements (Images used)

- https://towardsdatascience.com/how-bad-will-the-coronavirus-outbreak-get-predicting-the-outbreak-figures-f0b8e8b61991 (Outbreak Prediction)
- https://commons.wikimedia.org/wiki/File:Lasswell%E2%80%99s_Model_of_Communication.gif (Lasswell Model of Communication)
- https://stemlounge.com/content/images/2019/12/animated_algorithms.gif (Animated Algortihms)
- https://tenor.com/view/coronavirus-c%C3%B3mo-detener-al-coronavirus-how-to-stop-coronavirus-match-fire-gif-16603787 (Burning sticks)
- https://i.stack.imgur.com/qldty.gif (Forest Fire Model)
- https://digiphile.files.wordpress.com/2016/01/twitter-gid.gif?w=640 (Tweets Flying)
- https://docs.google.com/presentation/d/1TeiJSdd5UjT3YajNt3XkUA0HfK_EOjx14wcUvbAwm9c/ (Basics Slides)
- https://arxiv.org/pdf/1708.06401.pdf (Hawkes Process Demo Image)

Summary of Research Work

1. DNHP (Full Version). ***Choudhari J.,*** *Bhattacharya I., Dasgupta A., Bedathur S.* (To be Submitted to WSDM - 2021)
2. Unified MTPP. ***Choudhari J.,*** *Bhattacharya I., Dasgupta A., Bedathur S.* **TPP Workshop: NeurIPS-2019**
3. Discovering Topical Interactions in Text-Based Cascades Using HMHP. ***Choudhari J.,*** *Dasgupta A., Bhattacharya I.,Bedathur S.* **ICDM-2018**
4. Saving Critical Nodes with Firefighters is FPT. *Misra N.,* ***Choudhari J.,*** *Dasgupta A., Ramanujan M.,* **ICALP 2017**

1. Efficient Hierarchical Clustering for Classification and Anomaly Detection. Doshi I., Sajjala S., Bhatt R., Dasgupta A., **Choudhari J.**, (Aug.-2020 ArXiv)
2. One Pass Sampling for Symmetric Tensor Factorization. Shit S., Chhaya R., Dasgupta A. **Choudhari J.** (ICML 2020)
3. Nearly Optimal Space Efficient Algorithm for Depth First Search. Gupta M., Sharma S. **Choudhari J.** (2019 - ArXiv)
4. On Structural Parameterizations of Happy Coloring, Empire Coloring and Boxicity. Reddy V., **Choudhari J. WALCOM-2018**
5. BioGen: Automated Biography Generation. Ambavi H., Garg A., Garg A., Sharma M., Sharma R., **Choudhari J.**, Singh M. **JCDL-2019**
6. Contextual Emotion Detection Using Deep Learning. Pamnani A., Goel R., Singh M., **Choudhari J.**, **SEMEVAL-2019**
7. AgriBot: Agriculture-Specific Question Answer System. Jain N., Jain P., Kayal P, Sahit J. Pachpande S., **Choudhari J.**, Singh M. **STEM-2018**

1. Multiple Source Fault Tolerant Approximate Shortest Path. *Gupta M., Chugh K.,* **Choudhari J.** (TBD)
2. Eigenvector estimation in an Online setting. *Doshi I., Dasgupta A.* **Choudhari J.** (TBD)
3. Coresets for KL-Divergence (non-metric spaces). *Dasgupta A.,* **Choudhari J.**, Chhaya R. (TBD)
4. Approximately Counting Triangles in a graph in Random walk model. Haddadan S., Dasgupta A., **Choudhari J.** (TBD)
5. Estimating the number of nodes in a graph in Random walk model. Haddadan S., Dasgupta A., **Choudhari J.** (TBD)

Thank You!

Questions?

Acknowledgements

Prof. Anirban Dasgupta
Advisor

Dr. Indrajit Bhattacharya
Mentor

Prof. Neeldhara Misra
Collaborator

Prof. Chetan Pahlajani

DSC Members

FDC Members

Prof. Manoj Gupta
Collaborator

Prof. Balaraman Ravindran
External Examiner

Acknowledgements

Funding Institutions

TCS Research Fellowship

IARCS

Special Mention

Dr. Dinesh Garg

Prof. Bireswar Das

Prof. Krishna Kanti Dey

Collaborators & Friends

Acknowledgements

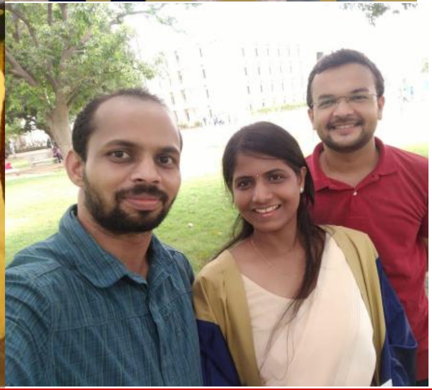
Friends

Acknowledgements

Friends



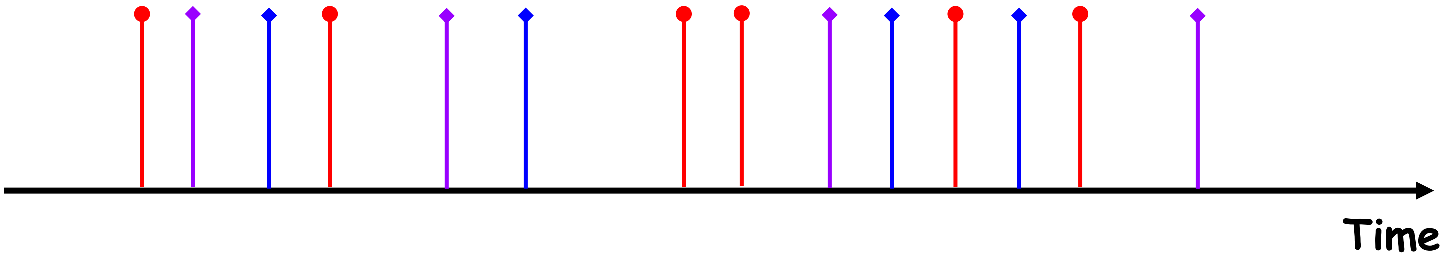


Acknowledgements

Family

Marked Temporal Point Process (MTPP)



$$\mathcal{H} = \{e_0 = (t_0, \eta_0), e_1 = (t_1, \eta_1), \ldots, e_n = (t_n, \eta_n)\}$$

$$t_i \in \mathbb{R}, \eta_i \in \mathbb{Z}$$

- Sequence of events of type η_i at times t_i
  - Continuous Time
  - Discrete, continuous (or mixed) marks (could be vector of marks)

Self-exciting Point Process (Univariate Hawkes Process)

Fig. 1

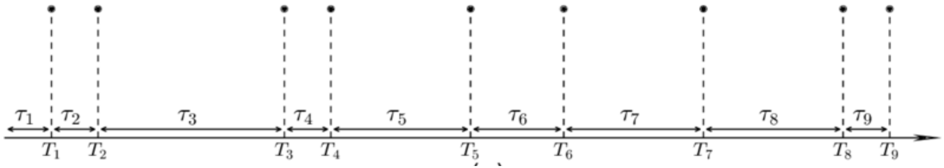


Fig. 2

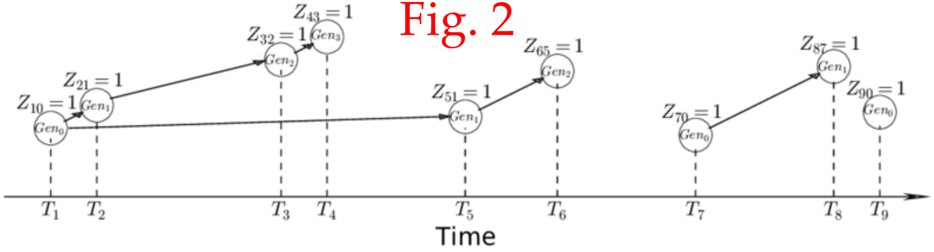
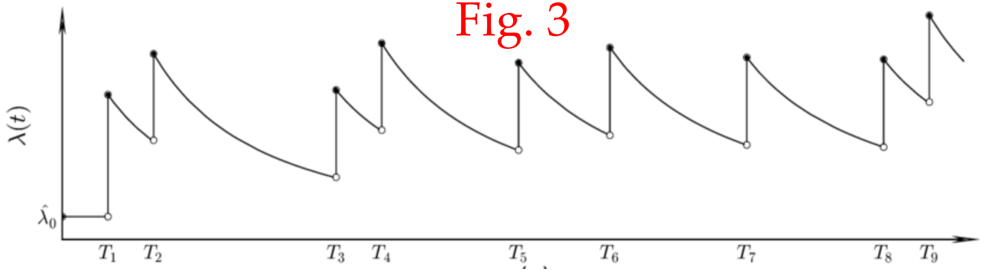


Fig. 3



Time-stamps are characterized by an intensity function:

$\lambda(t)dt := Pr\,(\text{event in } [t + dt) \,|\, \mathcal{H}_{t}\text{-})$

Let $c_n = v$

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

Dataset + Baseline

- **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)**

HMHP Results: *Semi-Synthetic Dataset*

PARENT IDENTIFICATION			
	HMHP	HWK+DIAG	HWKxLDA
ACCURACY	**0.58**	0.36	0.37
RECALL @1	**0.595**	0.373	0.380
RECALL @3	**0.778**	0.584	0.589
RECALL @5	**0.838**	0.674	0.678

Higher the better

NETWORK RECONSTRUCTION			
	HMHP	HWK+DIAG	HWKxLDA
MRE	**0.448**	0.565	0.552
MRE ($N_{uv} \geq 100$)	**0.398**	0.520	0.496

Lower the better

HMHP performs **better** at both **Parent Identification** and **Network Reconstruction** tasks.

DNHP Results: *Semi-Synthetic Dataset*

PARENT IDENTIFICATION		
	DNHP	HMHP
ACCURACY	0.45	0.28
RECALL @1	0.52	0.29
RECALL @3	0.73	0.46
RECALL @5	0.81	0.56

Higher the better

NETWORK RECONSTRUCTION		
	DNHP	HMHP
MRE	0.39	0.69
MRE ($N_{uv} \geq 100$)	0.28	0.66

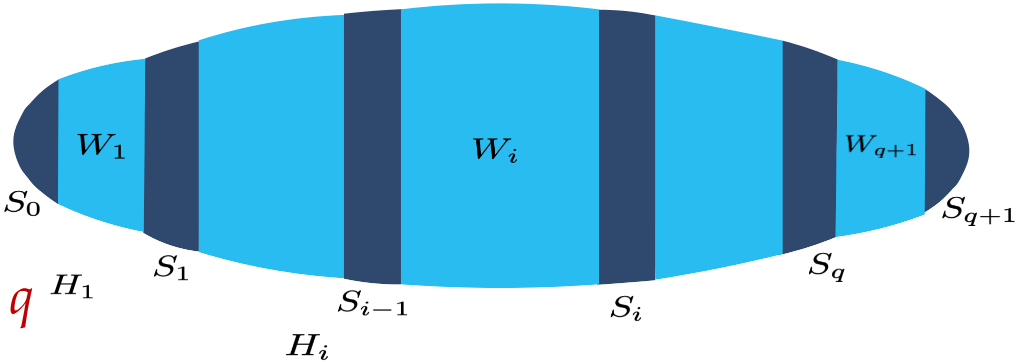
Lower the better

DNHP performs **better** than **HMHP** at both **Parent Identification** and **Network Reconstruction** tasks.

Tight Separator Sequence

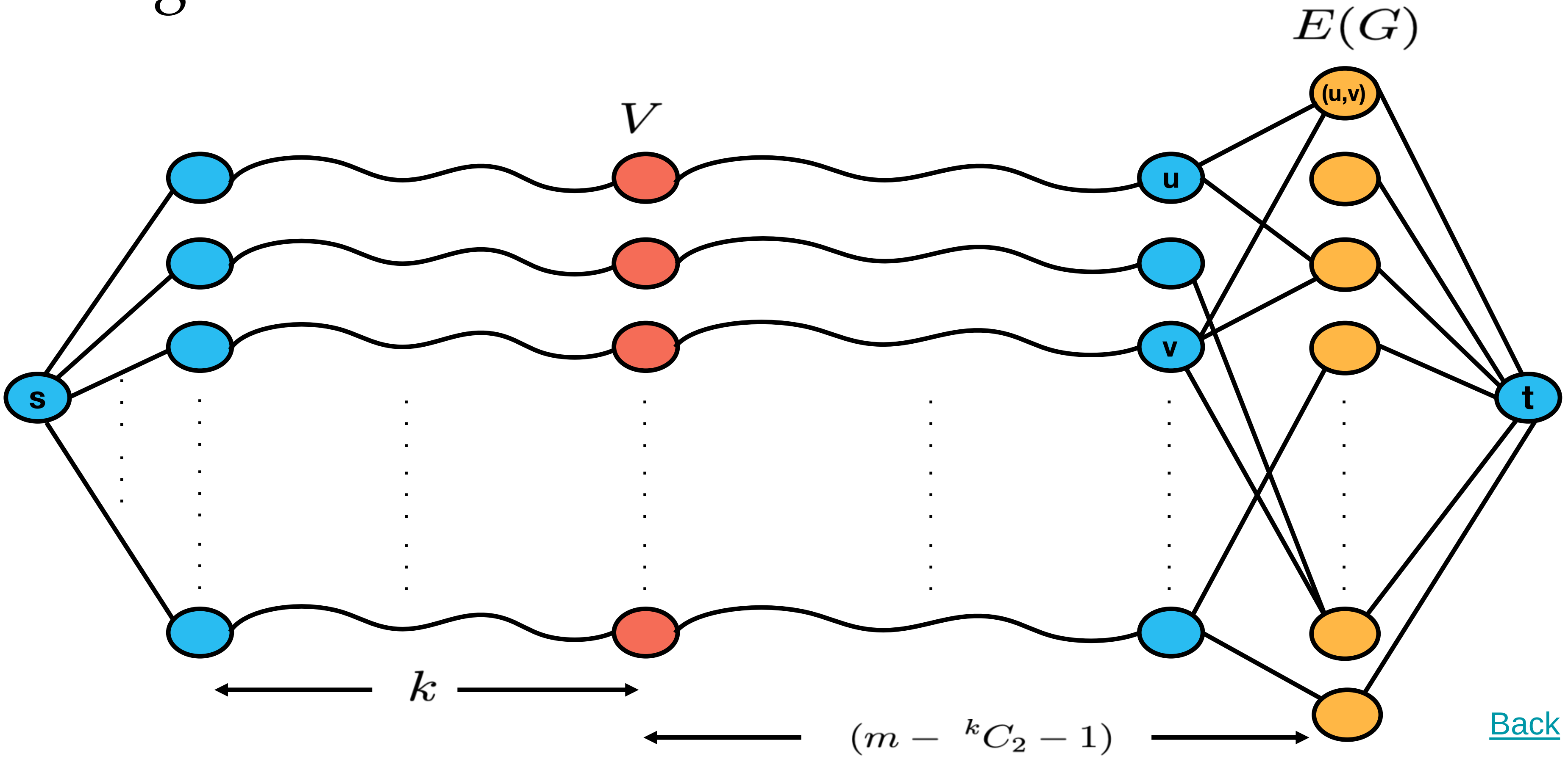
Let X, Y be two subset of vertices in the graph G.

Then, a ***tight (X,Y)-reachability sequence of order k*** *is an ordered collection* $\mathbf{H} = \{\mathbf{H_1}, \mathbf{H_2}, \ldots \mathbf{H_q}\}$ *of sets of V(G) satisfying the following properties*



1. $H_1 \subset H_2 \subset \cdots \subset H_q$
2. $|N(H_i)| \leq k$, *for all* i, $1 \leq i \leq q$
3. $S_i = N(H_i)$, *for all* i, $1 \leq i \leq q$ *is a minimal (X,Y)-separator in G*

Saving a Critical Set: Para-NPC



Back