CNeRG IIT Kharagpur

Analyzing the dynamics of information diffusion in social networks

Ayan Kumar Bhowmick (PhD Research Scholar)

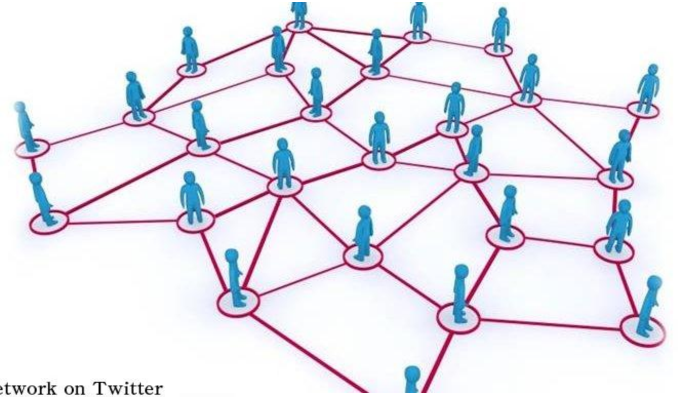
Advisor: Dr. Bivas Mitra

Department of CSE, Indian Institute of Technology Kharagpur, India

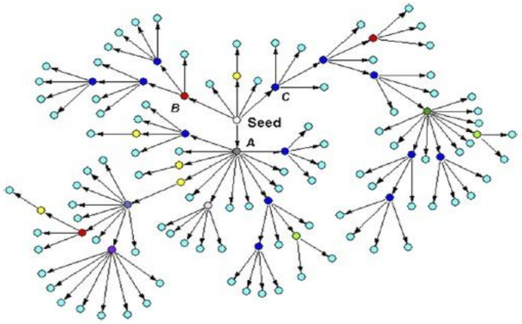
Talk at LIP6, 19th December, 2018

Information diffusion in social networks

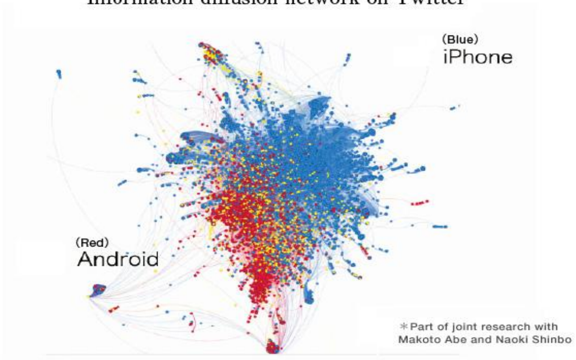
- Process by which a piece of information spreads among individuals through interactions



- Example:
    - Diffusion of topics in Twitter:

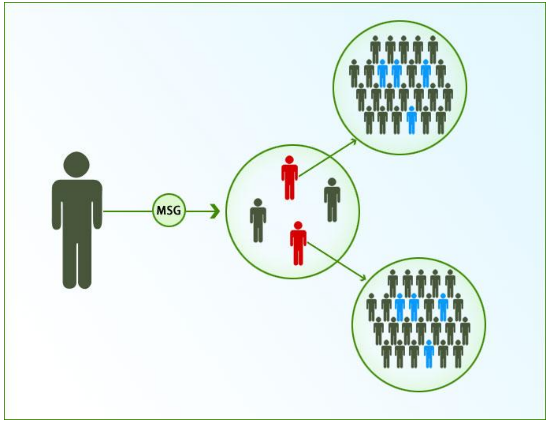


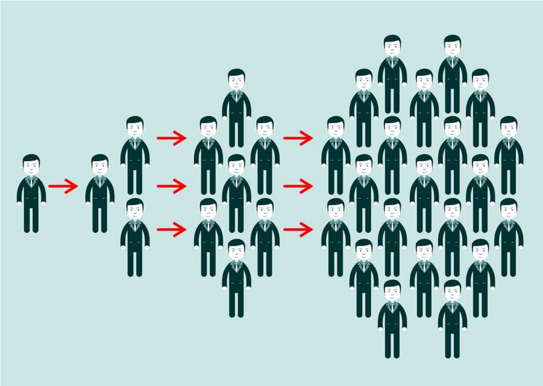
Information diffusion network on Twitter



Information diffusion in social networks

- Utility of information diffusion:
  - Identifying influential users
  - Content popularity
  - Identifying popular diffusion paths





Broad objectives

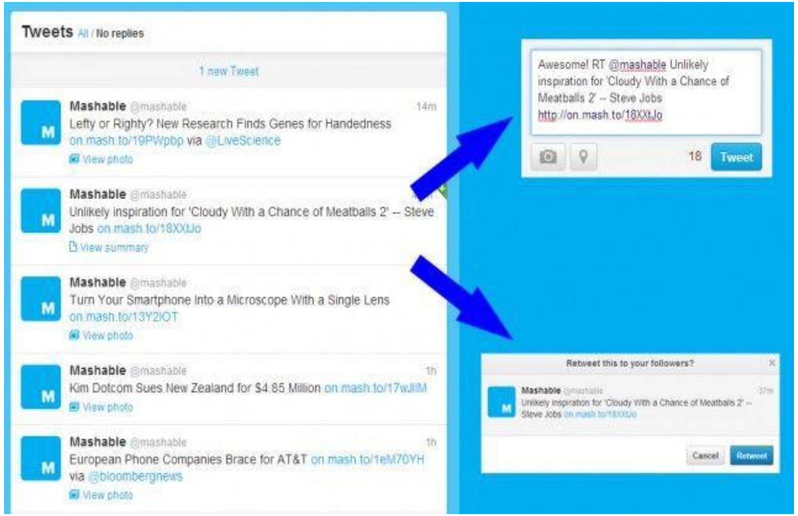
- Impact of information diffusion in online social networks:
  - Propagation of cascades in Twitter
- Model diffusion dynamics over follower network
  - Use signals from temporal data
    - Model transition of content to a new population
- Estimate morphology of cascades
  - Predict the influence tree structure of cascades solely from temporal signals

Background

Cascades in Twitter

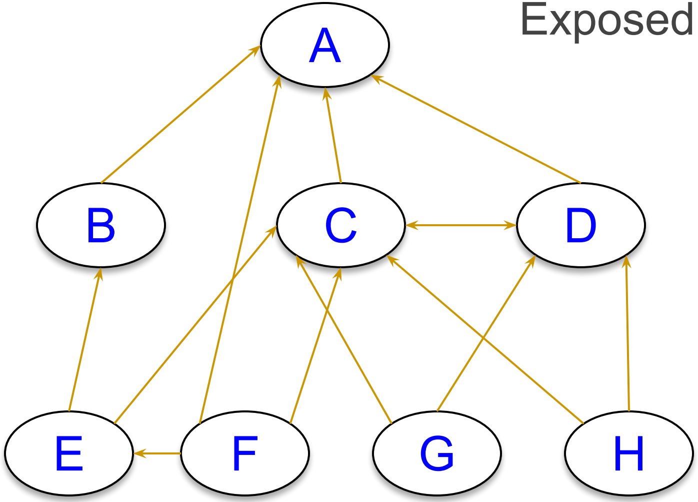
Users can reshare tweets posted by her followees through retweets

Long chain of retweets form cascades

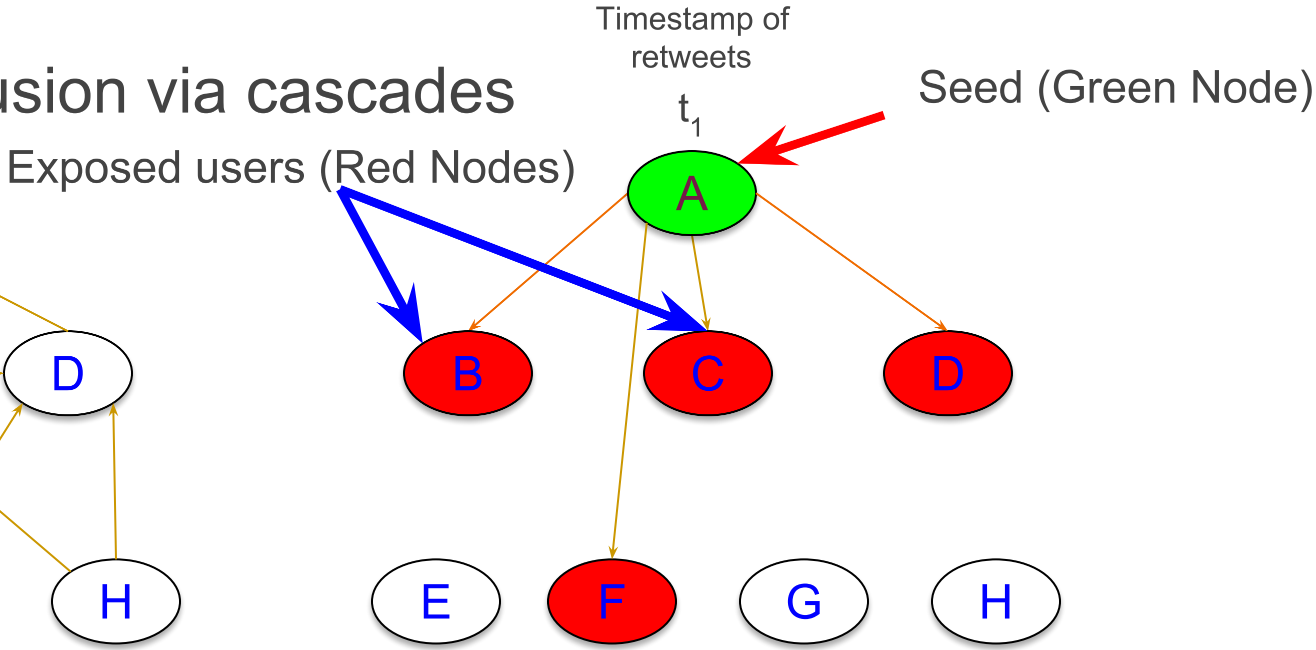


Background

Information Diffusion via cascades



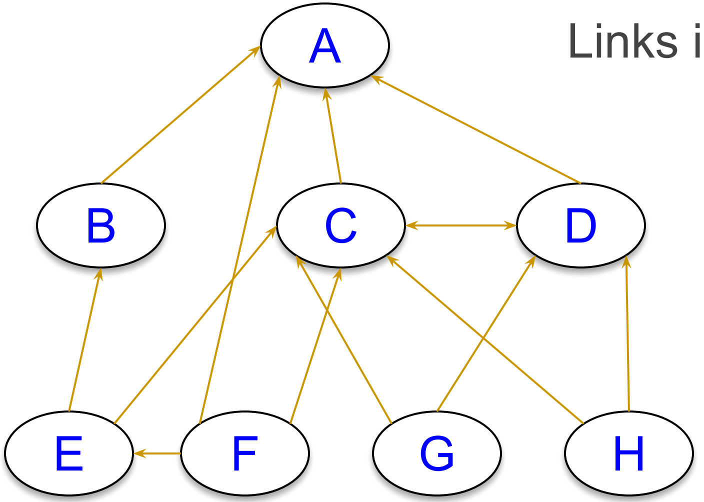
Twitter Follow Network



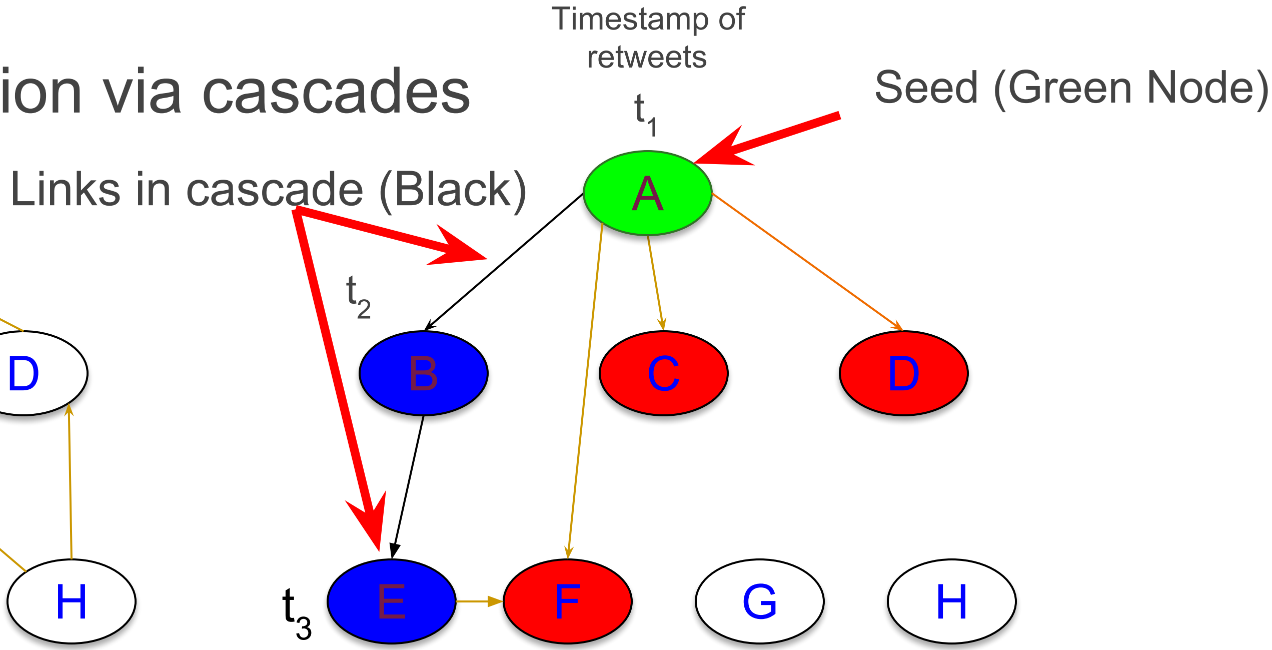
A growing cascade over the follower network

Background

Information Diffusion via cascades



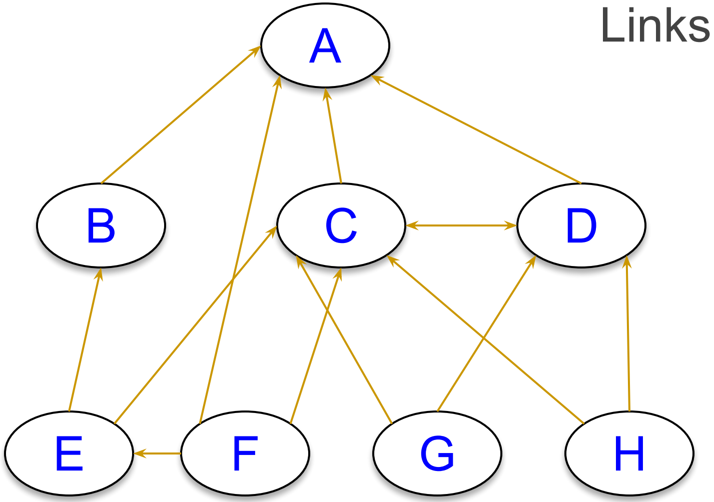
Twitter Follow Network



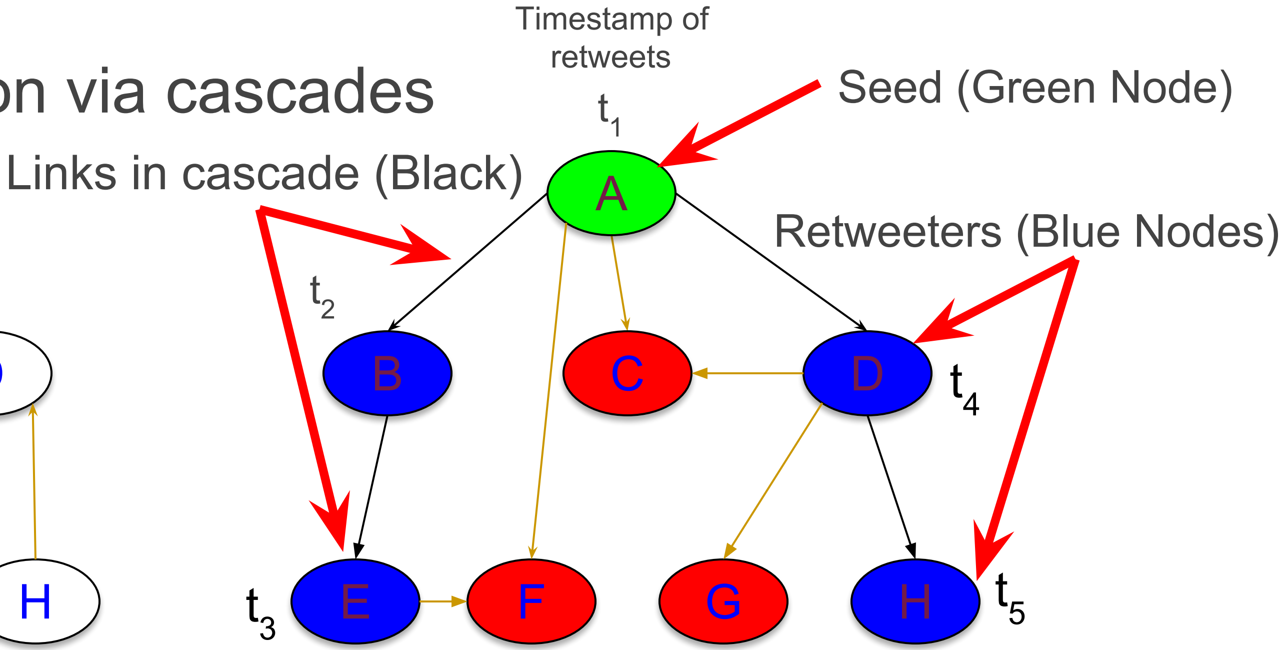
A growing cascade over the follower network

Background

Information Diffusion via cascades

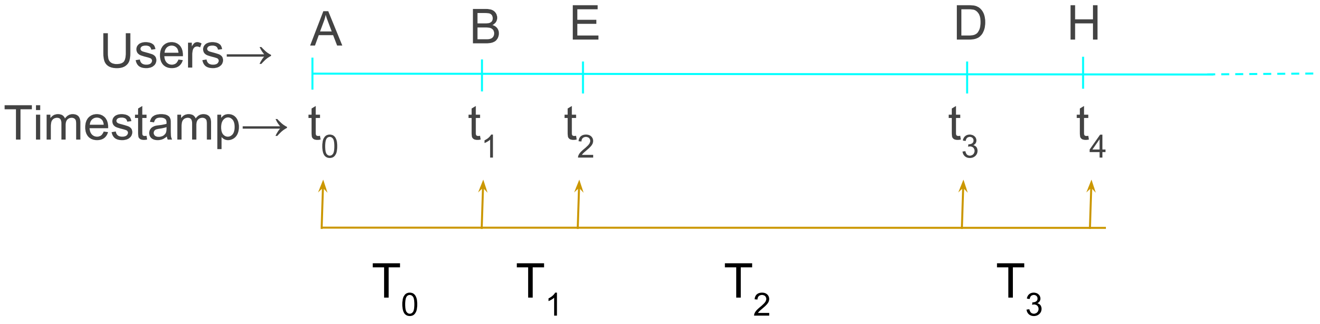


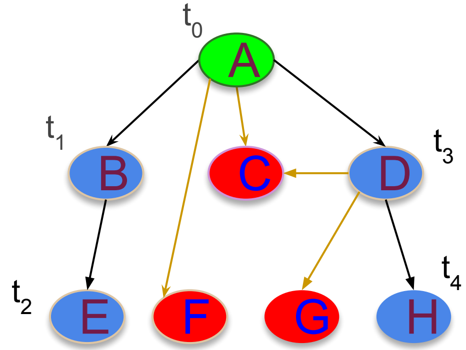
Twitter Follow Network



Influence Tree of a cascade

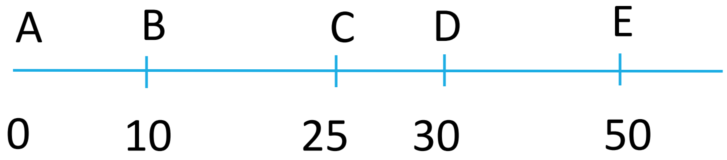
Background





- Let the time series for a cascade C of size n be denoted as $(t_0, t_1, \ldots, t_n)$
  - Pattern of inter-retweet time intervals for the cascade C : $T^C = (T_0, T_1, \ldots, T_{n-1})$
    - $T_i = t_{i+1} - t_i$ for i = 0,1,...,n-1

Example:



Corresponding inter-retweet intervals: **10** **15** **5** **20**

Related works

- Existing works on Twitter cascades:

Domain	
Modeling cascade popularity	Cheng et al. (WWW'14), Gao et al. (WSDM'15), Yang et al. (ICWSM'10), Cheng et al. (WWW'16)
Study of burstiness of cascades	Diao et al. (ACL'12), Myers et al. (WWW'14), Wang et al. (AAAI'15)
Using patterns of inter-retweet time intervals	Tavares et al. (PloS one'13), Webberley et al. (Comp Com'15), Ghosh et al. ('11)
Modeling the structure of cascades	Zong et al. (ICDM'12), Leskovec et al. ('06), Rodriguez et al. (WSDM'13)
Detection of structural holes	Lou et al. (WWW'13), Rezvani et al. (CIKM'15), Zhang et al. (ECMLPKDD'16)

- Limitations:
  - Identifying transitions between different regions of the network from temporal patterns ignored
  - Ignore signals from temporal data to predict influence tree structure of a cascade

Dataset

Algeria and Egypt Datasets

- Collection of tweets posted during events of the 2011 Arab Spring Movement
- Information available in the Dataset:
    1. Message_ID; Timestamp; User_ID; Content; Type: Tweet/Retweet
    2. Link to original Tweet id (for a retweet)
    2. Follower Network of users
- Dataset publicly available at:
  http://www.cnergres.iitkgp.ac.in/blog/2018/02/28/arab-spring-twitter-dataset/

Dataset	Tweets	Re-tweets	Cascades	Size of largest cascade	#Active Users
Algeria	65268	17269	5730	980	8814
Egypt	671417	188090	67539	432	13882

Dataset statistics

Temporal Pattern of (Re)tweets Reveal Cascade Migration

Ayan Kumar Bhowmick[1], Martin Gueuning[2,3], Jean-Charles Delvenne[3] & Renaud Lambiotte[2], Bivas Mitra[1]

1. Department of CSE, Indian Institute of Technology Kharagpur, India
2. naXys, University of Namur, Belgium
3. ICTEAM, Université Catholique de Louvain, Belgium

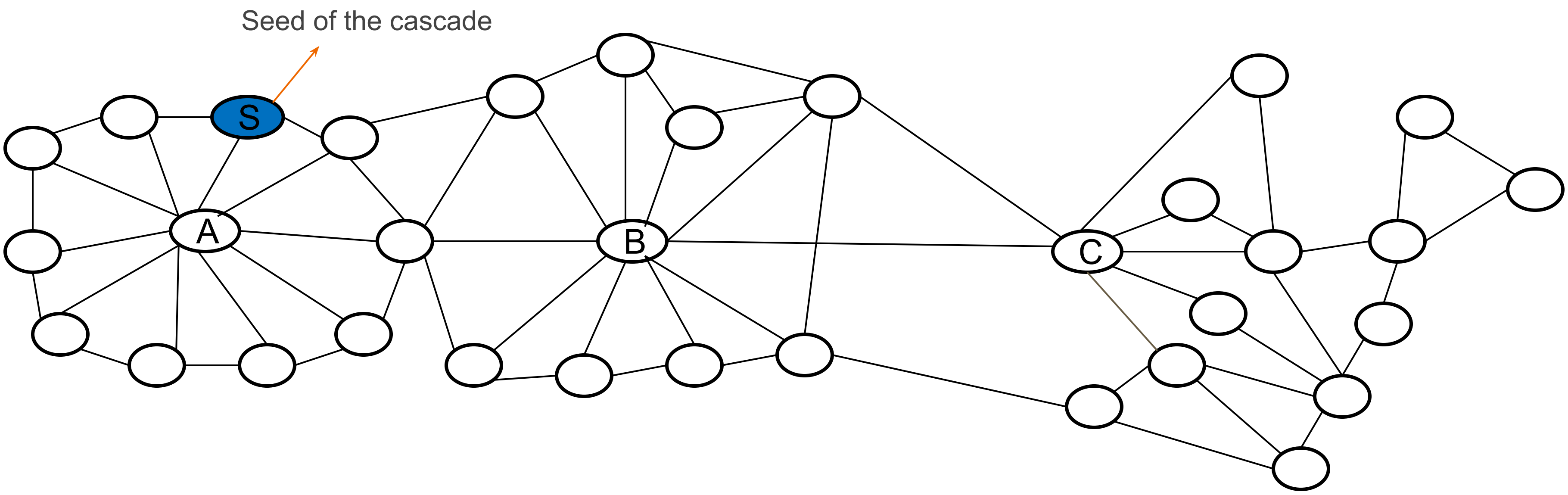
ASONAM 2017

Sydney, Australia

CNeRG IIT Kharagpur

Motivation: Diffusion of tweets

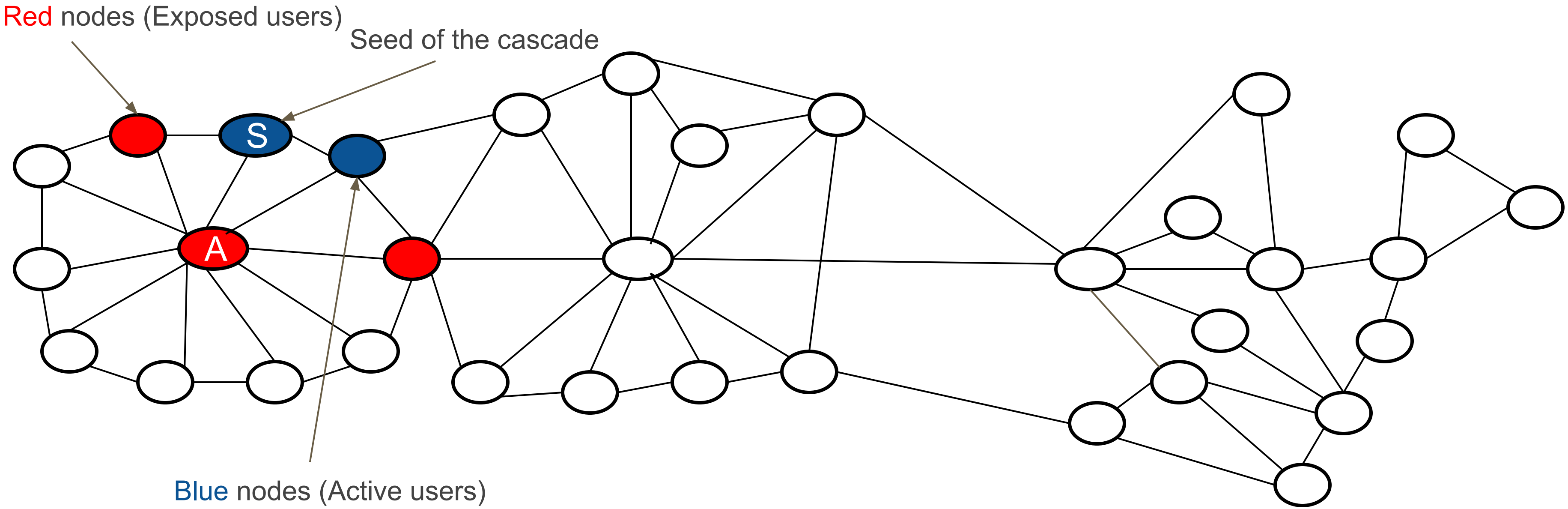
Consider propagating over a sample follower network



A sample follower network

Motivation: Diffusion of tweets

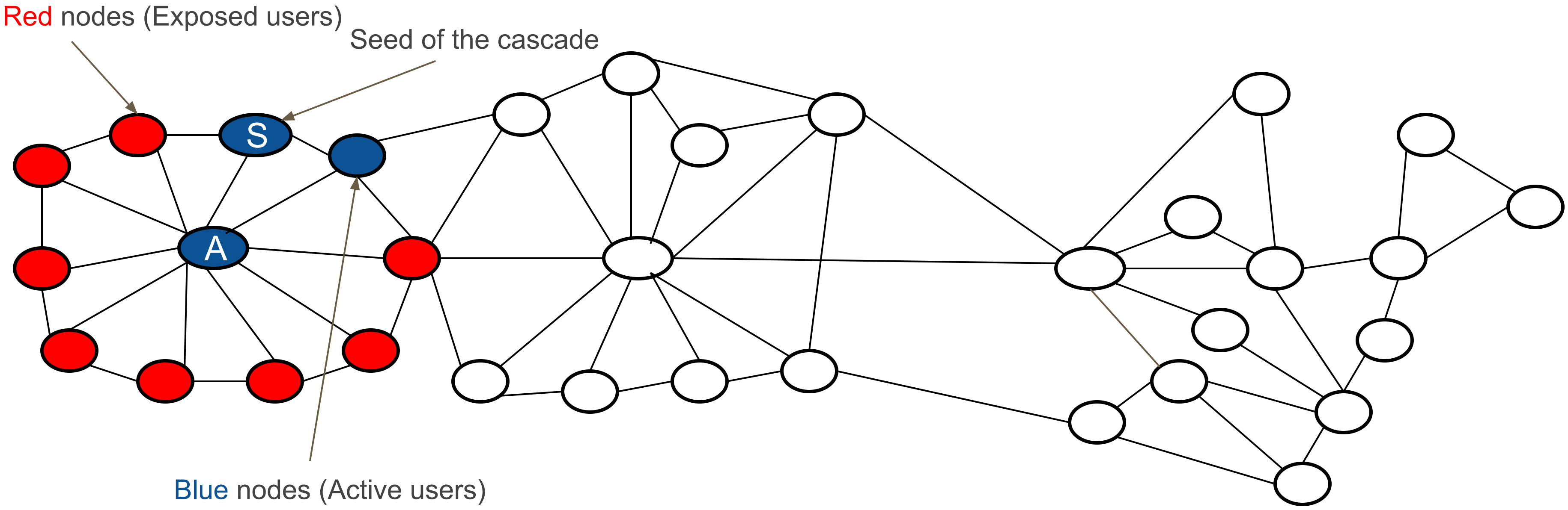
Tweets propagate over multiple localities



Propagation of Cascade over a sample follower network

Motivation: Diffusion of tweets

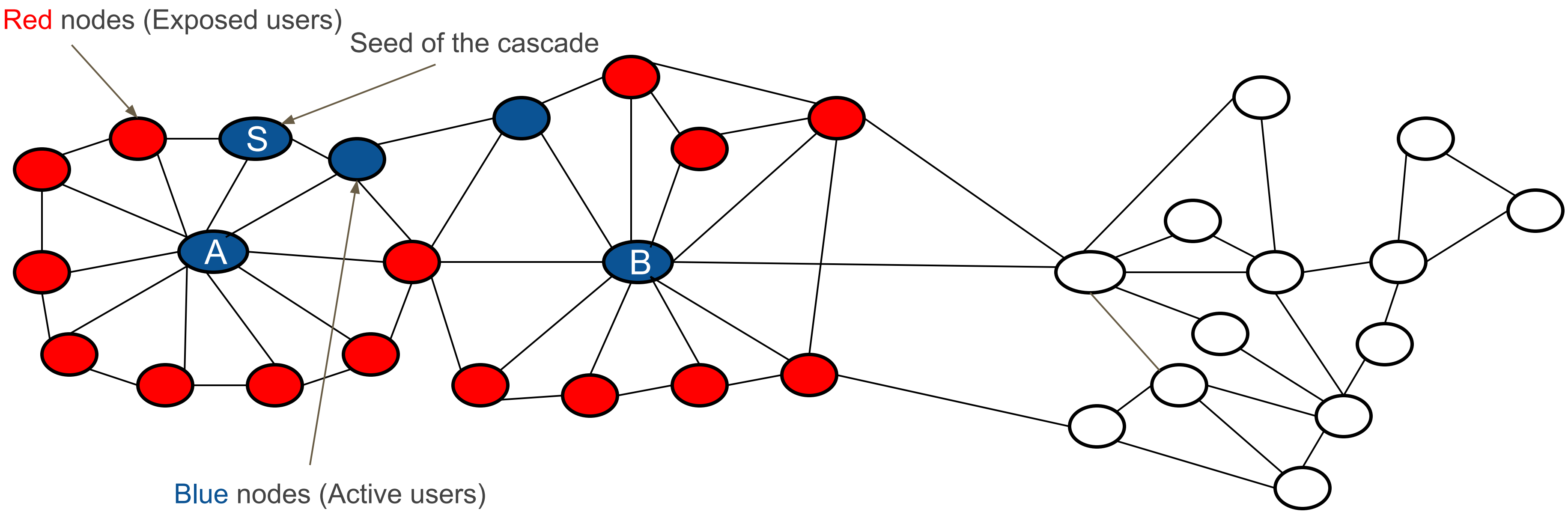
Tweets propagate over multiple localities



Propagation of Cascade over a sample follower network

Motivation: Diffusion of tweets

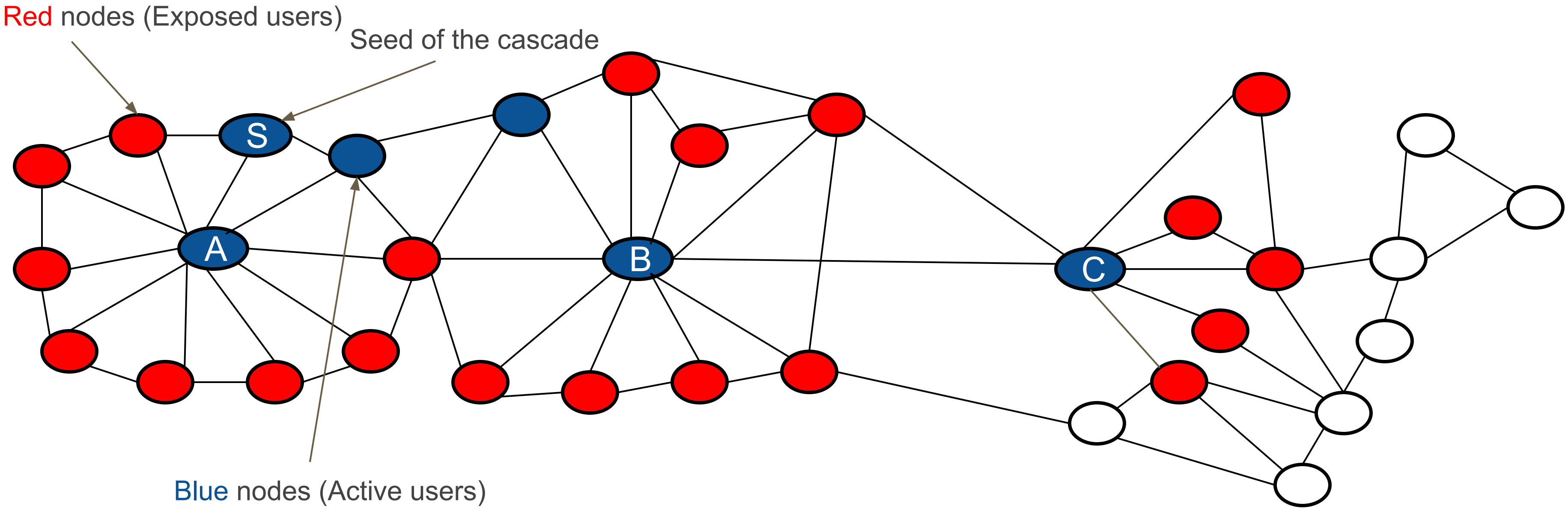
Tweets propagate over multiple localities



Propagation of Cascade over a sample follower network

Motivation: Diffusion of tweets

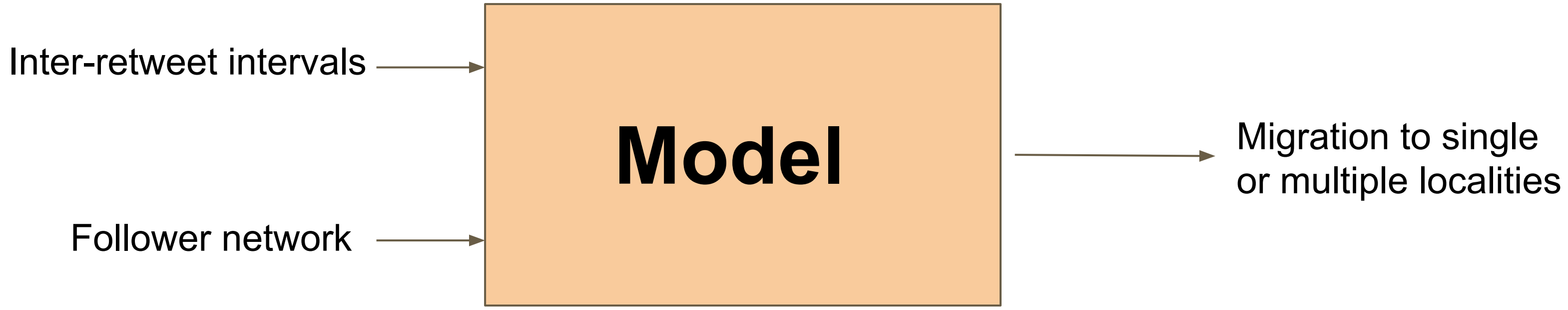
Tweets propagate over multiple localities



Propagation of Cascade over a sample follower network

Problem statement

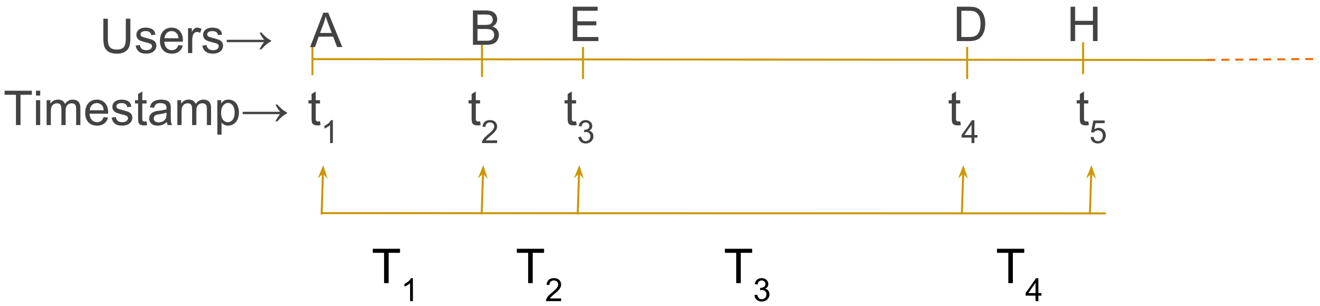
- Identify cascade diffusion over multiple localities (Detect migration)
    - Leveraging on the inter-retweet intervals $\mathbf{T_C}$

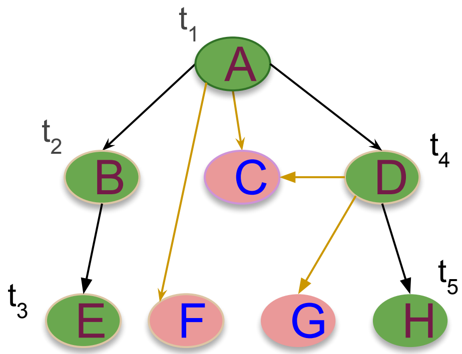


Outline

- **Inter-retweet intervals and Cascade diffusion**
- Empirical observations
- Analytical model
- Conclusion

Evolution of inter-retweet intervals





Let time series for a cascade C of size n be denoted as $(t_1, t_2, \ldots, t_n)$

Pattern of inter-retweet times for cascade C : $\mathbf{T_C = (T_1, T_2, \ldots, T_{n-1})}$

Mathematically, $\mathbf{i}^{th}$ inter-retweet time is calculated as:

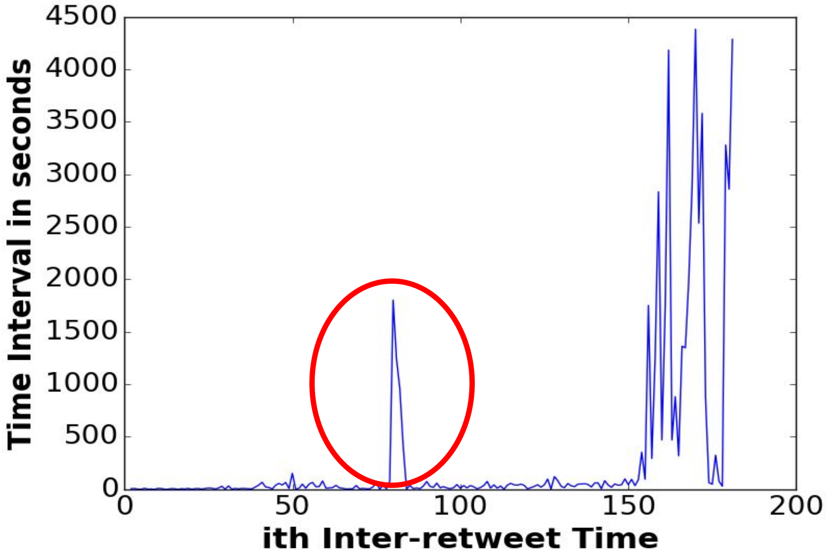
$\mathbf{T_i = t_{i+1} - t_i}$ for $i = 1,2,\ldots,n-1$

Example: Sequence of retweet timestamps in a cascade: 20 80 150 180 220 2000 2020

Corresponding inter-retweet intervals: **60 70 30 40 1780 20**

Evolution of inter-retweet intervals

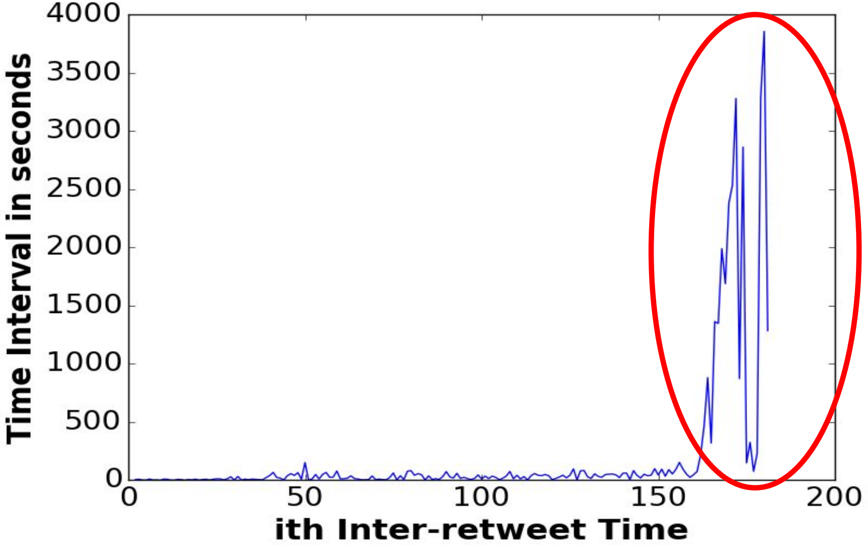
Classification of cascades based on pattern of inter-retweet intervals:



Type I (Early peak)

Peak at intermediate stage (40% retweets occurred)

80% of cascades for Algeria; 61% for Egypt



Type II (Late peak)

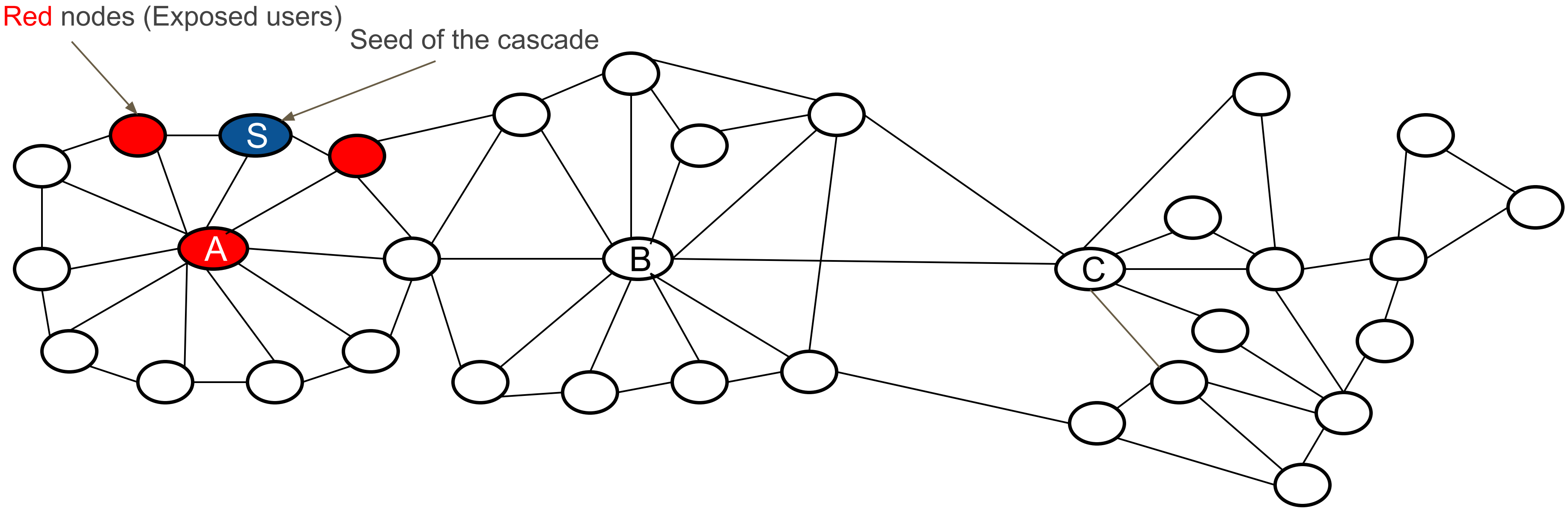
Peak at the end (more than 80% retweets occurred)

20% of cascades for Algeria; 39% for Egypt

Cascade diffusion

Introducing the concept of **diffusion locality** with respect to cascade diffusion

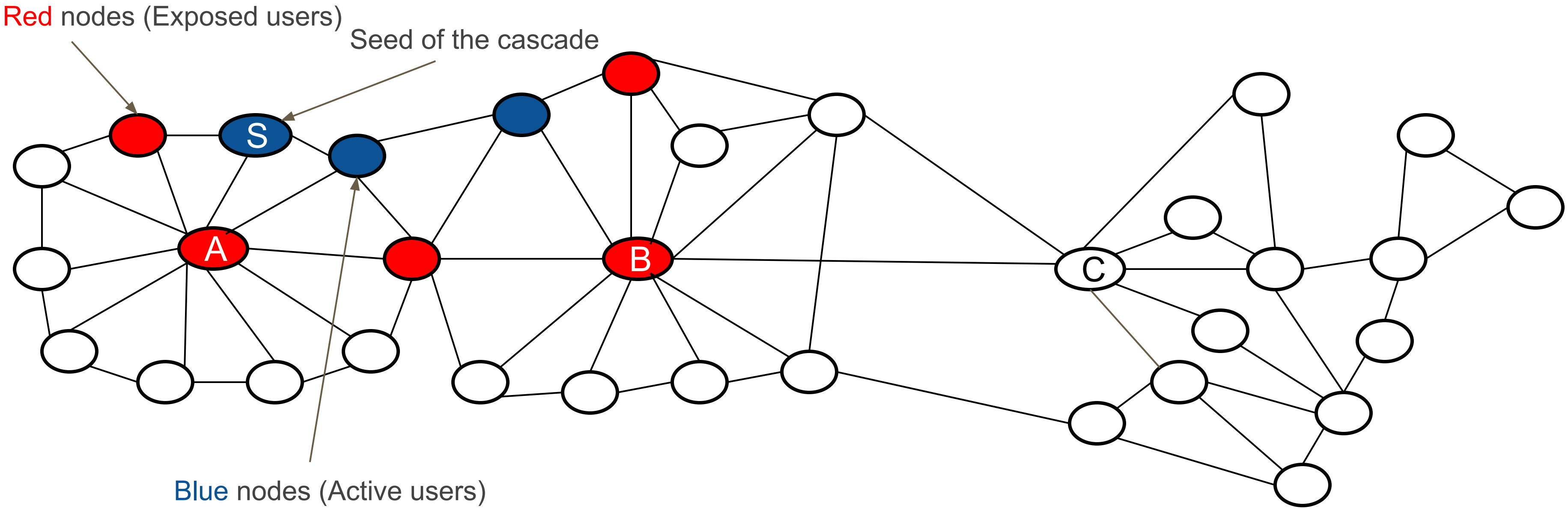
- **Diffusion locality: After each retweet event, the set of exposed users grows slowly**



Cascade diffusion over the sample follower network

Cascade diffusion

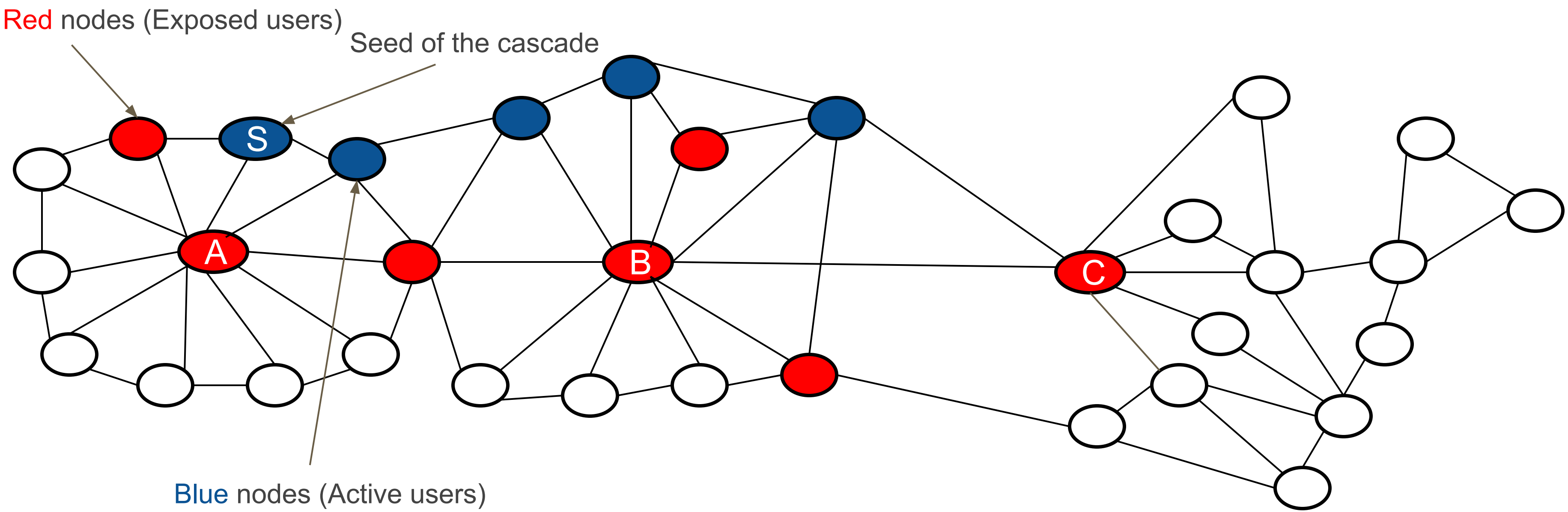
Diffusion locality grows incrementally with each retweet activity



Cascade diffusion over sample follower network

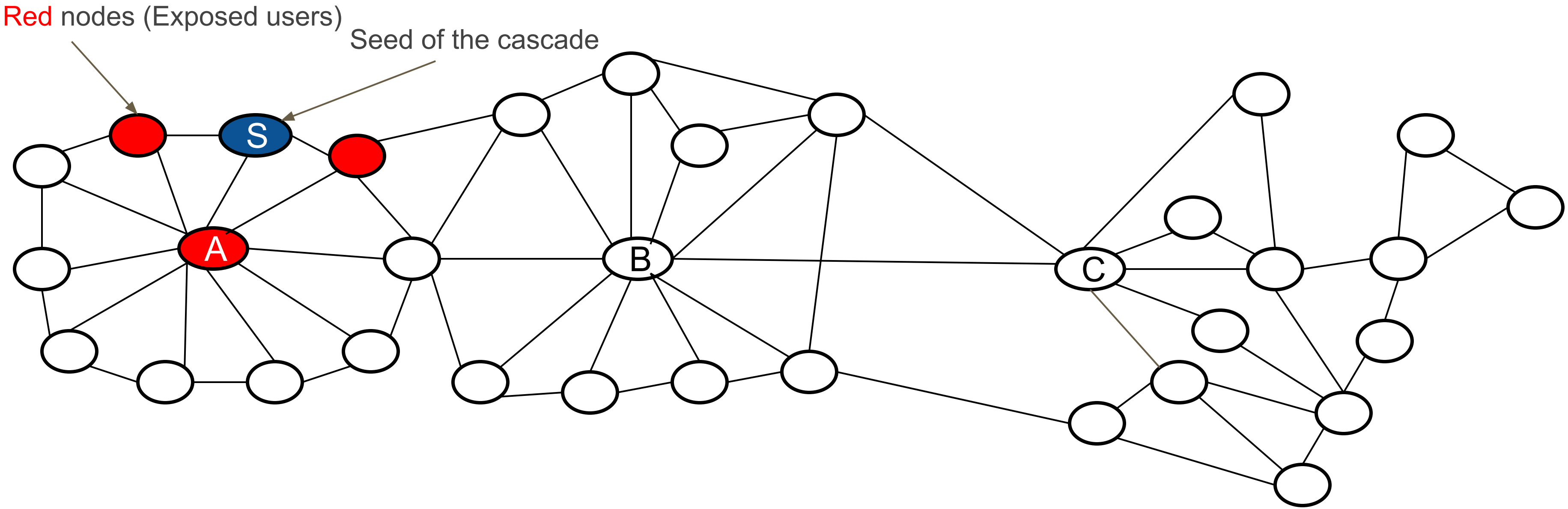
Cascade diffusion

Cascade confined to single **diffusion locality** due to small incremental growth in exposed set **(accretion)**



Cascade diffusion over sample follower network

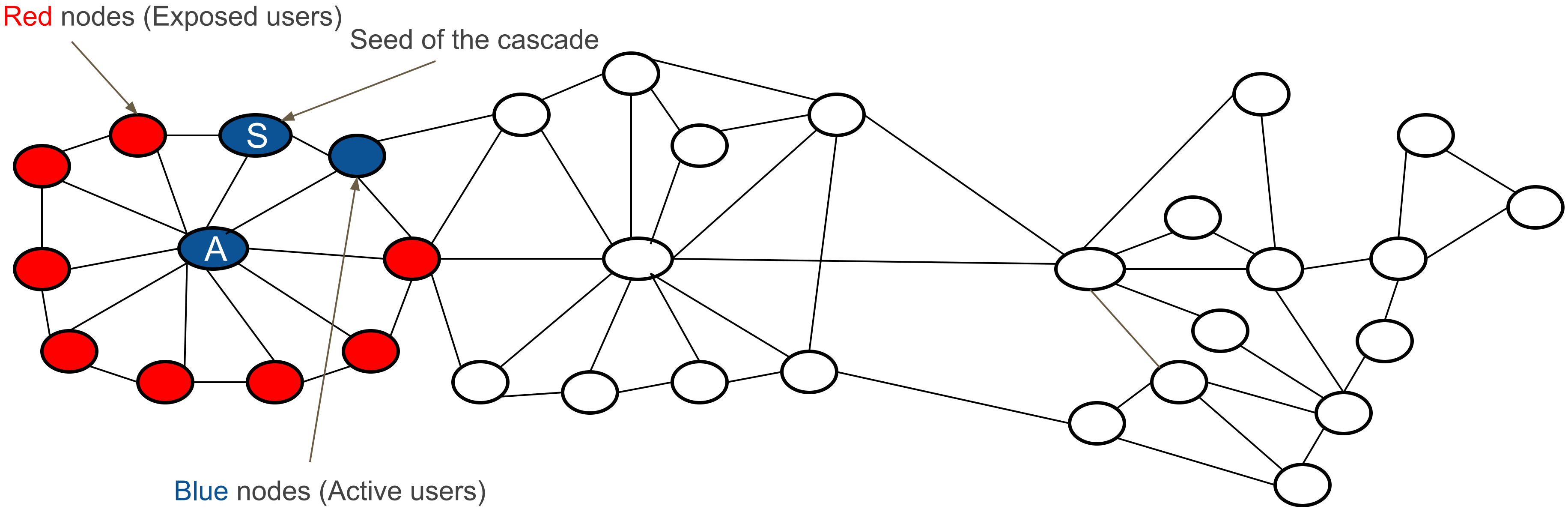
Cascade diffusion: Flush effect



Cascade diffusion over the sample follower network

Cascade diffusion: Flush effect

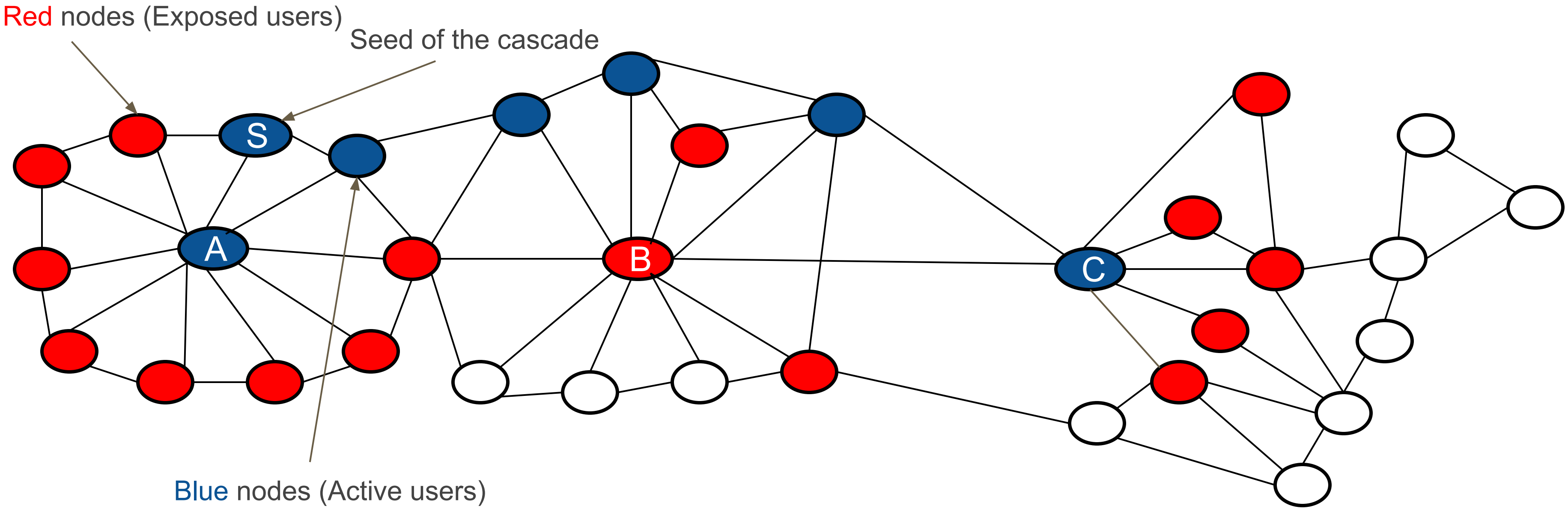
A single retweet by either A or C exposes cascade to a new **diffusion locality (migration of cascade)**



Cascade diffusion over sample follower network

Cascade diffusion: Flush effect

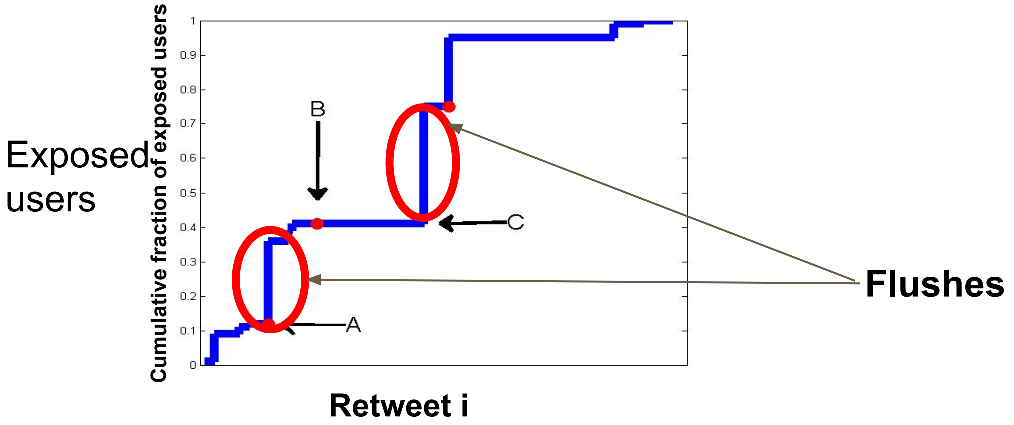
A single retweet by either A or C exposes cascade to a new **diffusion locality (migration of cascade)**



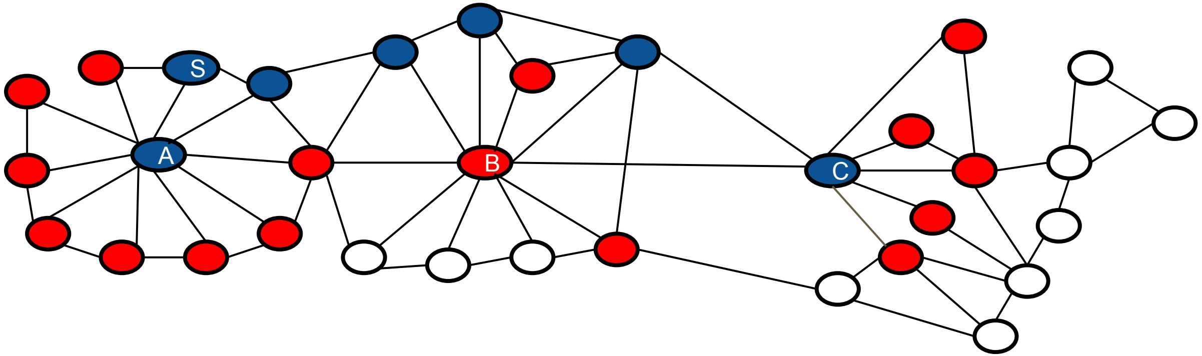
Cascade diffusion over sample follower network

Cascade diffusion : Flush effect

- Cascade migrates to a new diffusion locality due to retweets by **pivotal users** (A and C in our case)



Number of flushes

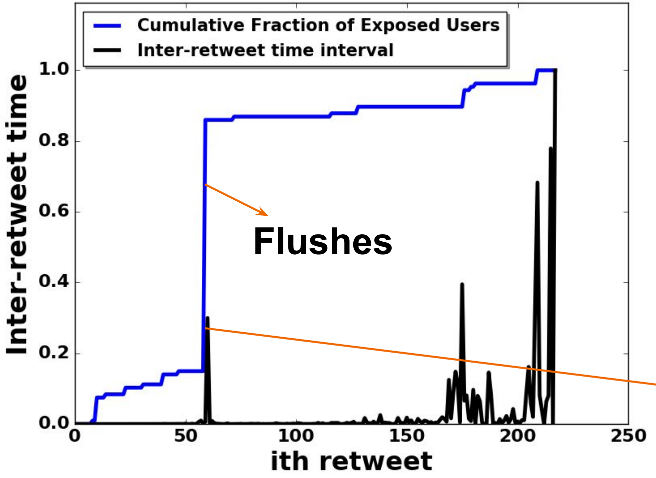


Outline

- Inter-retweet intervals and Cascade diffusion
- **Empirical observations**
- Analytical model
- Conclusion

Empirical observation (1)

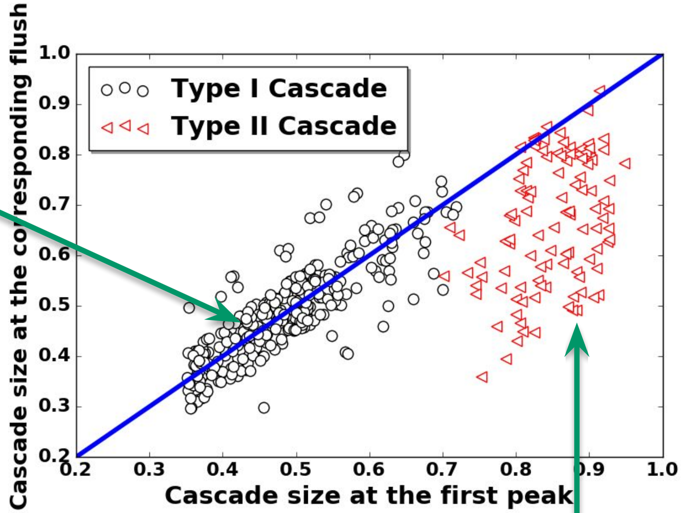
- Co-occurrence of flush effect and first peaks (Type I cascade)
  - Flushes and first peaks in Type I cascades occur close to each other
    - Co-occurrence observed for 82% of Type I cascades
    - No co-occurrence for Type II cascades (absence of early peak)



Saturation of the current locality

The occurrence of the first peak and the corresponding flush co-occur for a typical **Type I cascade**

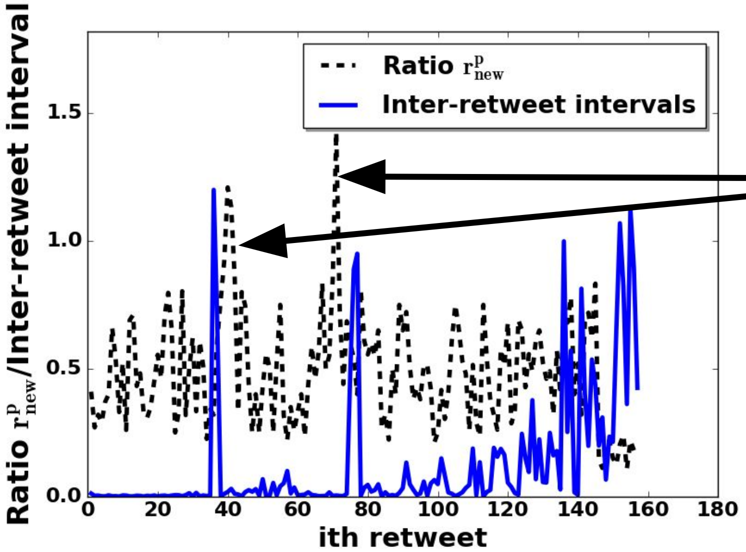
first peak and flush occur close to each other in Type I cascades;



No such correlation for Type II cascades

Empirical observation (2)

- Cascade migrates to new locality after first peak in Type I cascades
  - Ratio r_{new}^{p} of fraction of **newly exposed users after the peak** to the fraction **exposed prior to the peak** >1 for a retweet near a peak
    - Fraction newly exposed after the peak more than fraction exposed before the peak for Type I cascades
    - This ratio is greater than 1 for 72% of Type I cascades
    - Ratio is always <1 for all Type II cascades indicating no migration



r_{new}^{p}>1 for retweets near peaks at early stages of a typical Type I cascade; signify migration to new diffusion locality

Outline

- Inter-retweet intervals and Cascade diffusion
- Empirical observations
- **Analytical model**
- Conclusion

Analytical model

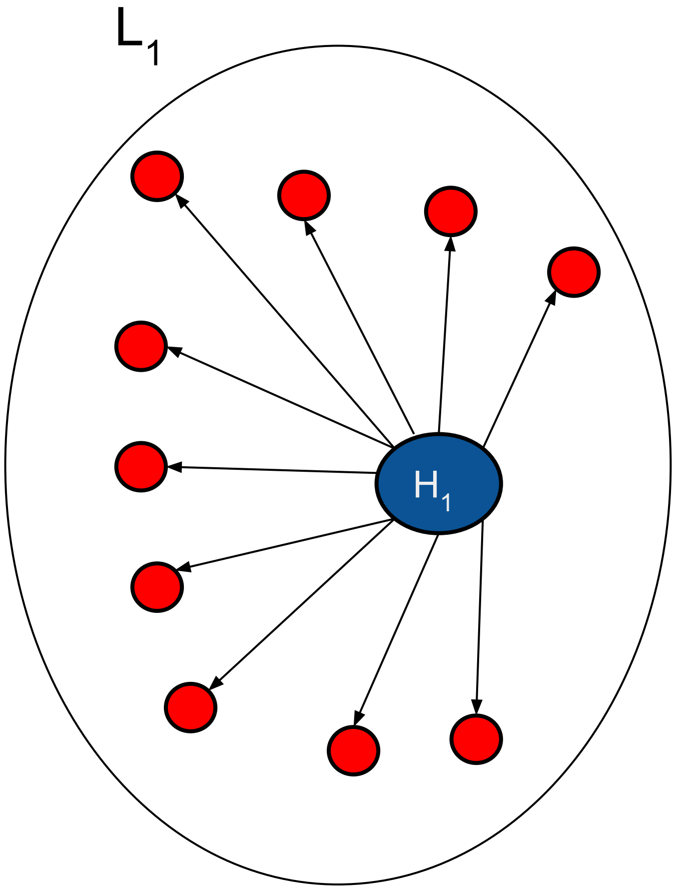
- Explain the empirical observations:
  - presence/absence of early peaks in inter-retweet intervals
  - co-occurence of peaks and saturation of current locality

- Model the cascade diffusion in two different cases
  - Case 1: Diffusion of the cascade in a single locality
  - Case 2: Cascade migration across multiple localities

Analytical model

Case 1: Diffusion in a single locality

- Diffusion locality L_1 approximated by a central node H_1
    - H_1 with v neighbors who retweet
    - v retweet times: $(X_1,\ldots,X_v) \sim f$
    - Time series of cascade: $(X_{(1)},\ldots,X_{(v)})$



Analytical model

Case 1: Diffusion in a single locality

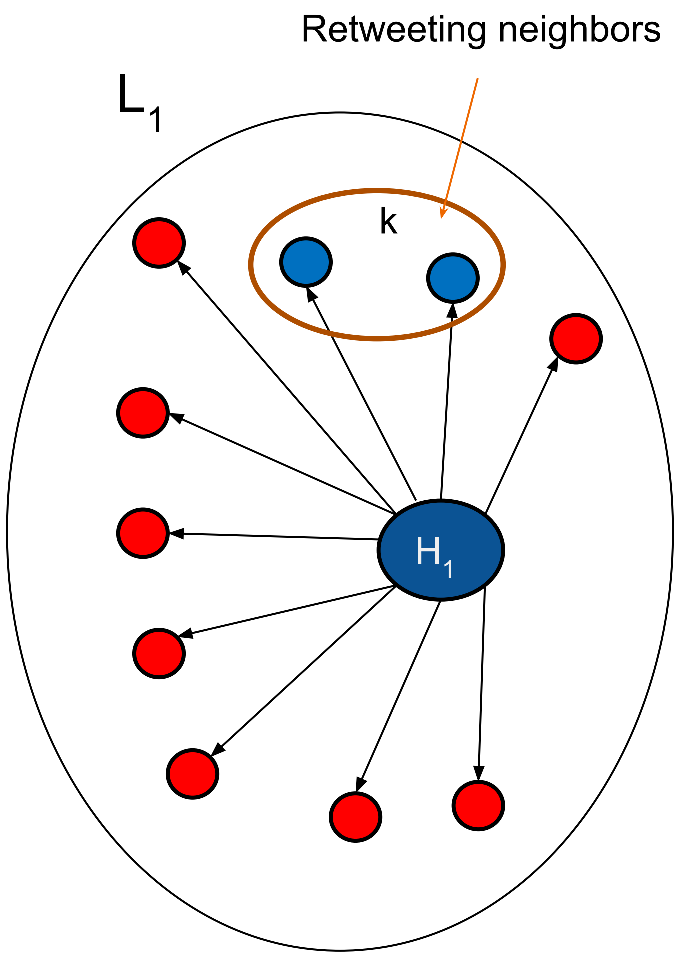
- Diffusion locality L_1 approximated by a central node H_1
  - H_1 with v neighbors who retweet
  - v retweet times: $(X_1,\ldots,X_v) \sim f$
  - Time series of cascade: $(X_{(1)},\ldots,X_{(v)})$
  - k: Number of retweeting (active) neighbors of H_1 at time t

Time series of cascade $(\underbrace{X_{(1)},\ldots,X_{(k)}}_{},\ldots\ldots\ldots X_{(v)})$

Time instance of retweet of user k, $X_{(k)}$ follows the distribution f_k

$$f_k(t) = \frac{\nu!}{(\nu-k)!\,(k-1)!}\, f(t)\, F(t)^{k-1}(1-F(t))^{\nu-k},$$

$$F(t) \equiv \int_0^t f(\tau)\,\mathrm{d}\tau \text{ is the cumulative function.}$$



Analytical model

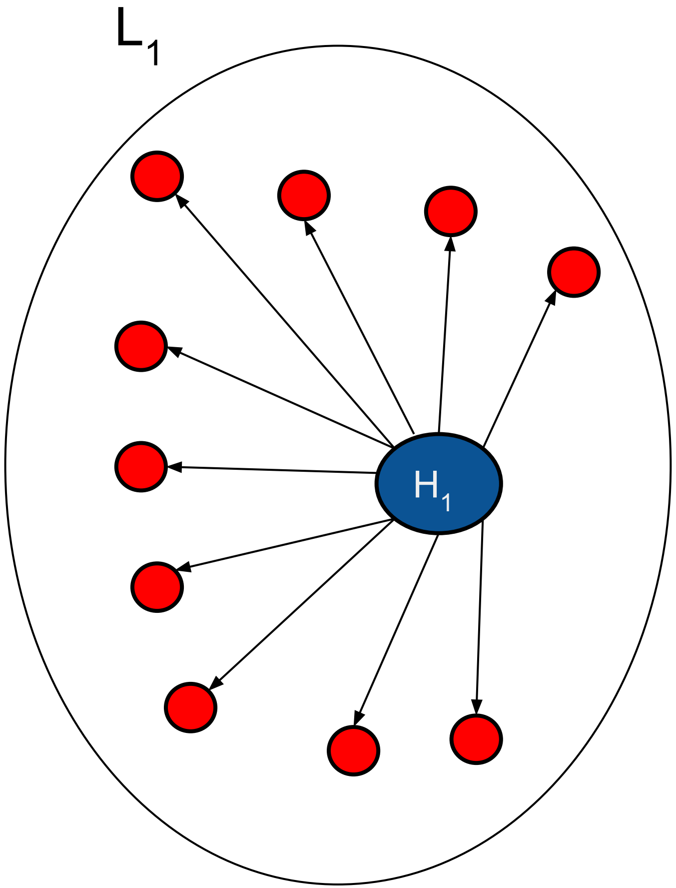
Case 1: Diffusion in a single locality

- Diffusion locality L_1 approximated by a central node H_1
  - H_1 with v neighbors who retweet
  - v retweet times: $(X_1,...,X_v) \sim f$
  - Time series of cascade: $(X_{(1)},...,X_{(v)})$
  - k →Number of active neighbors of H_1
- E_k: k^{th} inter-retweet interval of a cascade:

$$\langle E_k \rangle = X_{(k)} - X_{(k-1)}$$

$$\langle E_k \rangle = \int_0^{+\infty} t f(t) \frac{\nu!}{(\nu - k + 1)!\,(k-1)!} \cdots F(t)^{k-2}\,(1 - F(t))^{\nu - k}\,\nu \left[F(t) + \frac{1-k}{\nu} \right] \mathrm{d}t \qquad (2)$$

Value of E_k →Depends on f,k,v



Analytical model

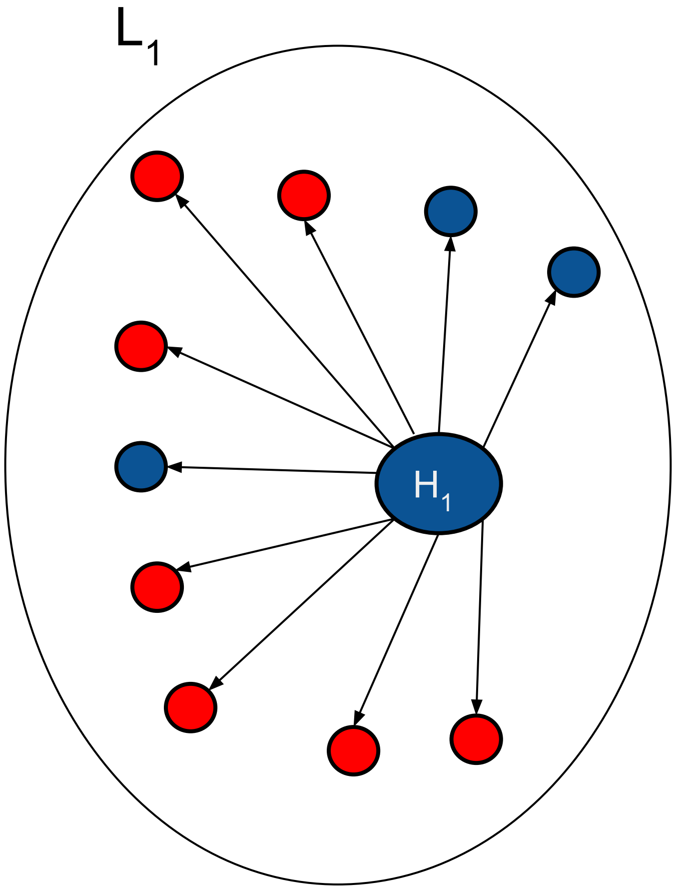
Case 1: Diffusion in a single locality

- Diffusion locality L_1 approximated by a central node H_1
  - H_1 with v neighbors who retweet
  - v retweet times: $(X_1,\ldots,X_v) \sim f$
  - Time series of cascade: $(X_{(1)},\ldots,X_{(v)})$
  - k →Number of active neighbors of H_1
- k^{th} inter-retweet interval of a cascade:

$$\langle E_k \rangle = \frac{1}{\lambda} \frac{1}{\nu - k + 1},$$

Simple case in which each retweet is a Markovian process, where

$$f \sim \mathrm{Exp}(- \lambda)$$



Analytical model

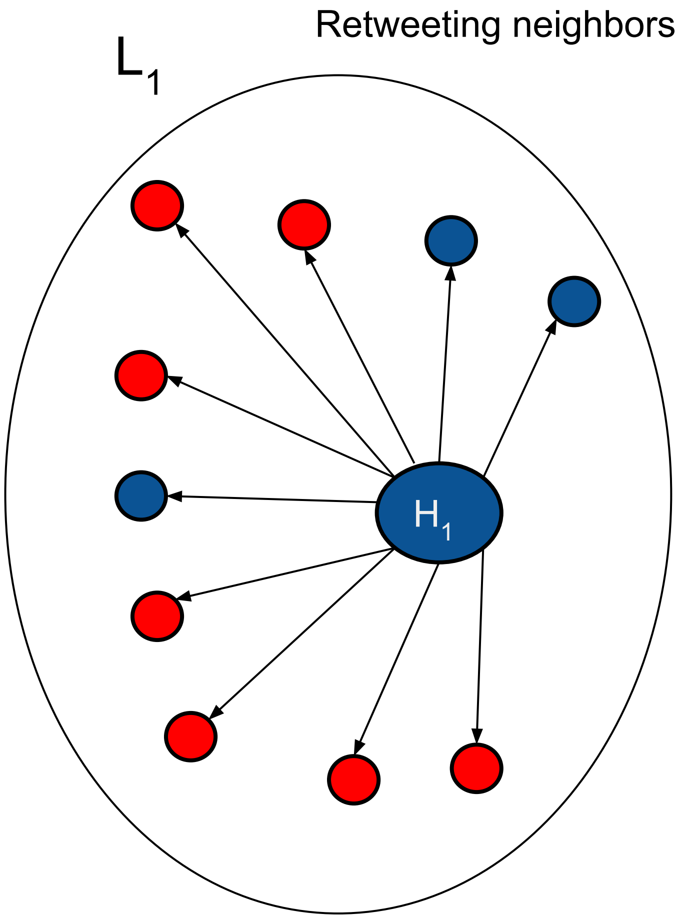
Case 1: Diffusion in a single locality

- Diffusion locality L_1 approximated by a central node H_1
  - H_1 with v neighbors who retweet
  - v retweet times: $(X_1,...,X_v) \sim f$
  - Time series of cascade: $(X_{(1)},...,X_{(v)})$
  - k →Number of active neighbors of H_1
- k^{th} inter-retweet time of a cascade:

$<E_k>$ →0 when k << O(v)

Simple case in which each retweet is a Markovian process, where

$f \sim \mathrm{Exp}(-\lambda)$



Analytical model

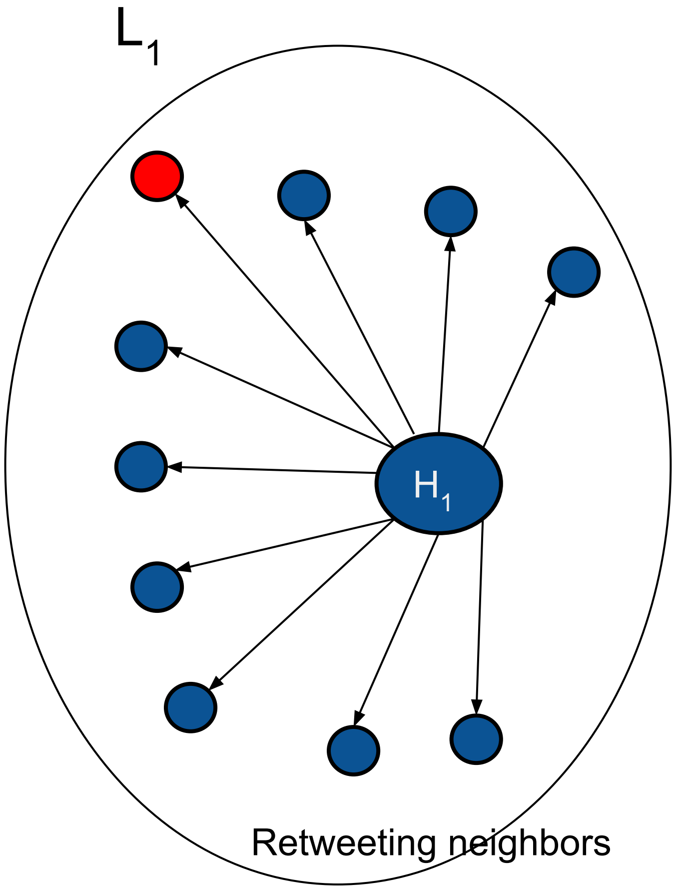
Case 1: Diffusion in a single locality

- Diffusion locality L_1 approximated by a central node H_1
  - H_1 with v neighbors who retweet
  - v retweet times: $(X_1,\ldots,X_v) \sim f$
  - Time series of cascade: $(X_{(1)},\ldots,X_{(v)})$
  - k →Number of active neighbors of H_1
- k^{th} inter-retweet time of a cascade:

$$\langle E_k \rangle > 1/\lambda \text{ when } k \rightarrow O(v)$$

Simple case in which each retweet is a Markovian process, where

$$f \sim \text{Exp}(-\lambda)$$



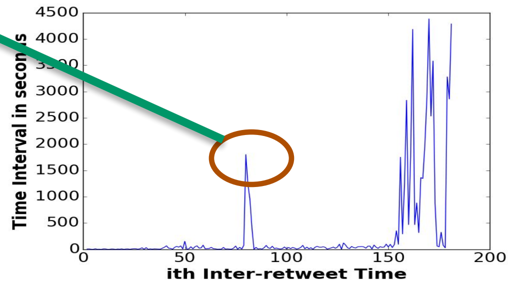
Validation of Analytical model

Key take away from the model

Inter-retweet time intervals are low (Ek=0) when the diffusion locality is not saturated (k << O(v))
Inter-retweet time intervals are high (Ek > 0) as it approaches saturation (k = O(v))

Low inter retweet interval: less than 20% of this interval can be considered as 'low'
High inter retweet interval : More than 80% is considered as 'high'.

Reference point



Observation:
Egypt dataset (for all cascades)
1. When saturation level (k/v) is lower than 0.3, almost 94% of corresponding **inter-retweet intervals exhibit low value**.
2. On the contrary, when **saturation level is greater than 0.8**, 89% of the corresponding **inter-retweet intervals exhibit high value**.

Conclusion

- Detected cascade transition across multiple diffusion localities from temporal pattern of retweets
  - Introduced the concept of **diffusion locality** specific to a cascade
  - Identified different types of cascades based on presence/absence of temporal peaks

- Built an analytical model that explains
  - Co-occurence of first peaks and migration of Type I cascades
  - Inter-retweet intervals <Ek> provide signatures of the **content saturation** in the current diffusion locality
  - Peak in the inter-retweet interval (manifested by the rise and fall in <E_k>) indicates the **migration of the cascade** to a new diffusion locality after saturation of the current locality.

- Validated the analytical model from empirical data
  - Detected co-occurence of first peaks and subsequent migration to new locality with good accuracy
  - Identified correlation between locality saturation and corresponding inter-retweet interval

Constructing Influence Trees from Temporal Sequence of Retweets: An Analytical Approach

Ayan Kumar Bhowmick, G.Sai Bharath Chandra, Yogesh Singh, Bivas Mitra

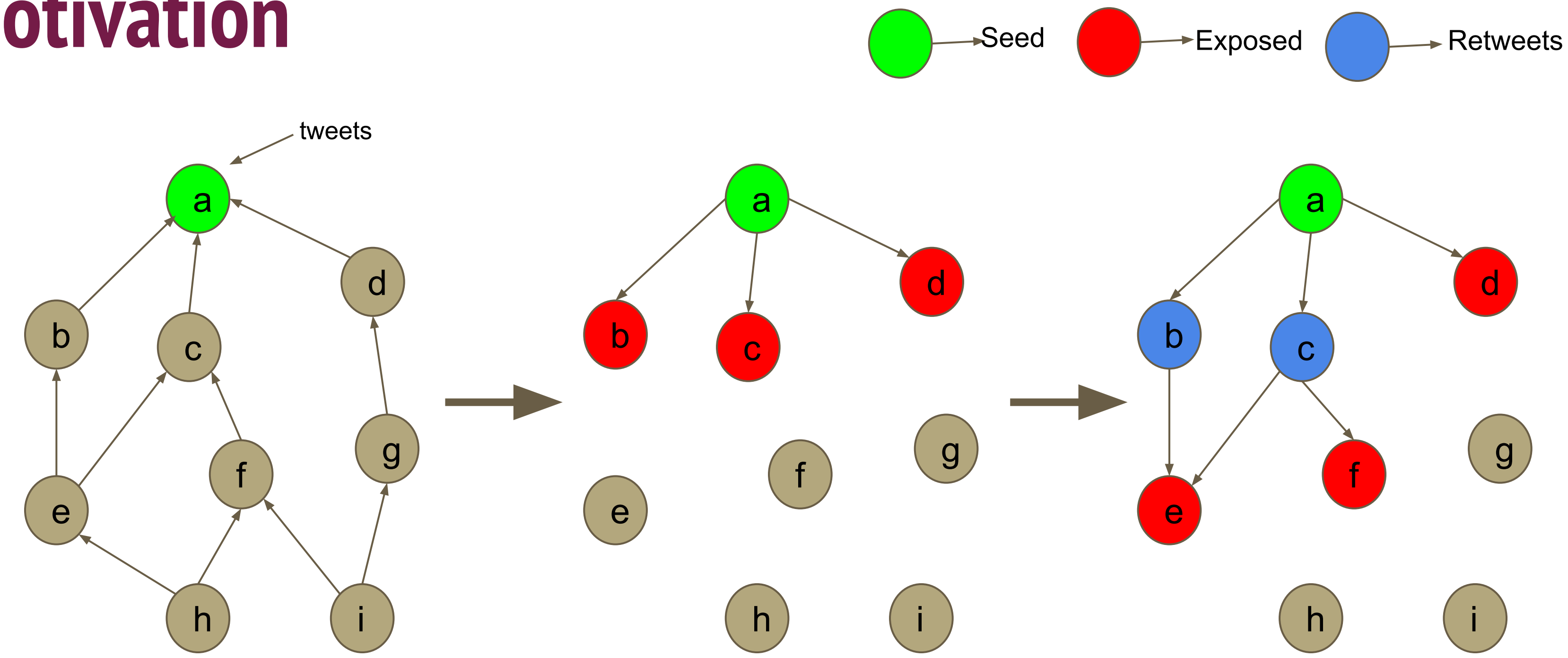
Department of CSE, Indian Institute of Technology Kharagpur, India

IEEE BigData 2018

Seattle, USA

CNeRG IIT Kharagpur

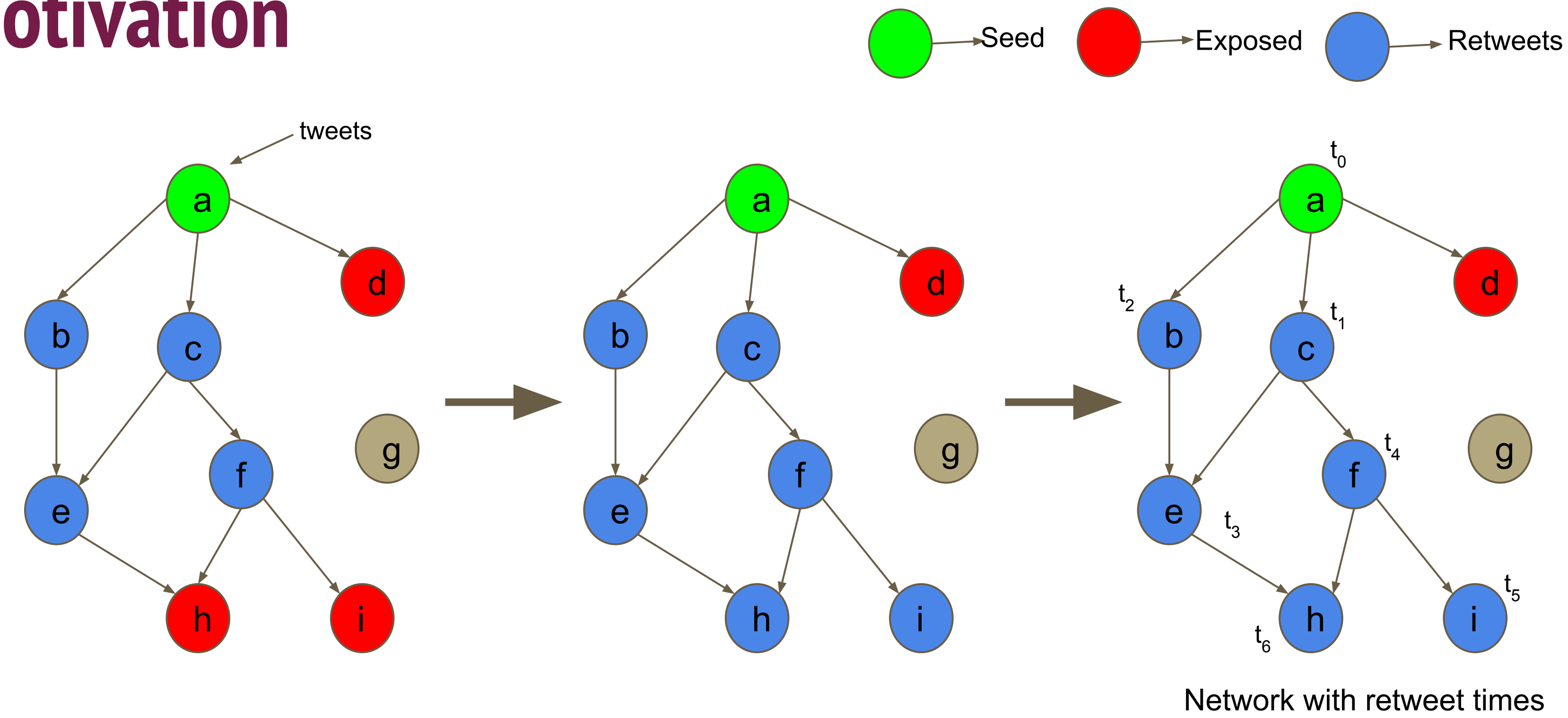
Motivation



Twitter follower network

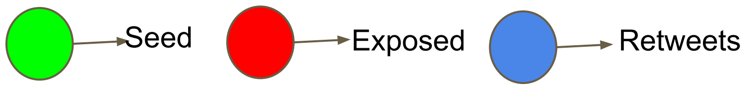
A growing cascade reaching more and more users through retweets

Motivation

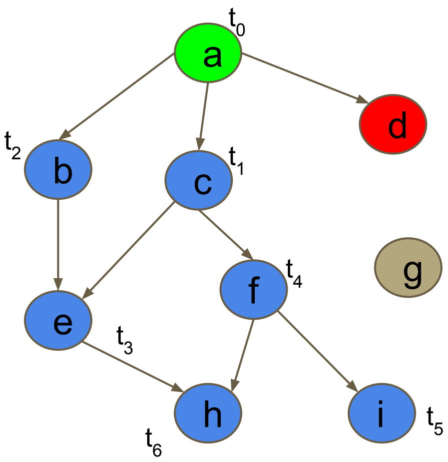


Chain of retweets of a tweet form a cascade

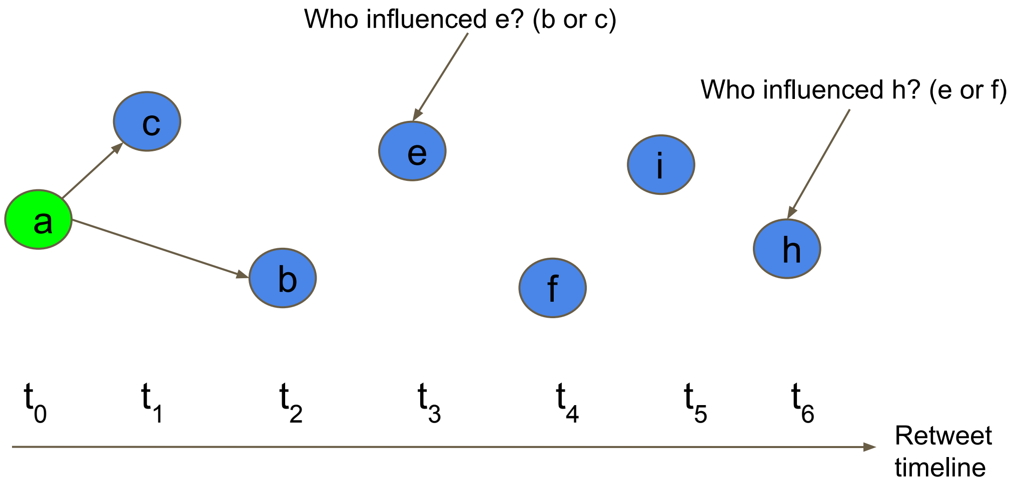
Motivation



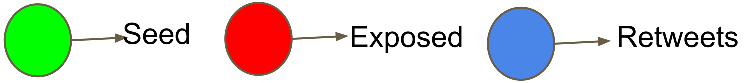
- Influence tree : A directed tree denoting who-influenced-whom relationship
- Node = retweeted user
- Edge (A -> B) = A Influences B

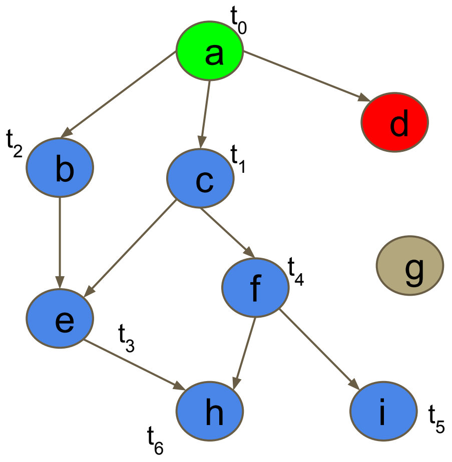


Network with retweet times

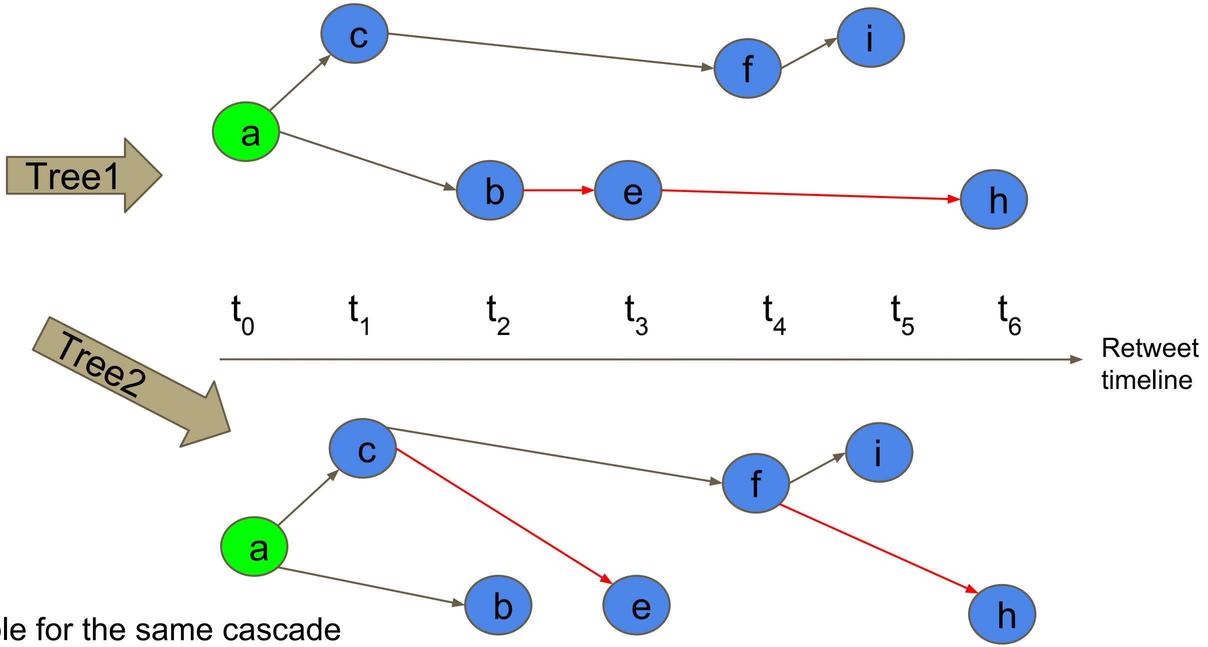


Challenges





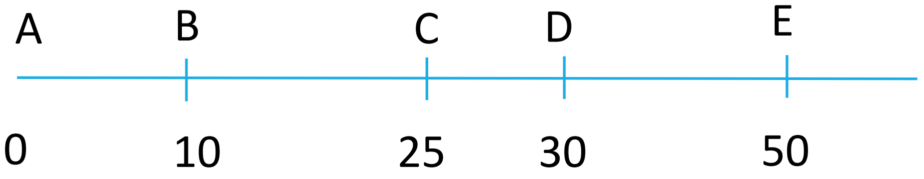
Network with retweet times



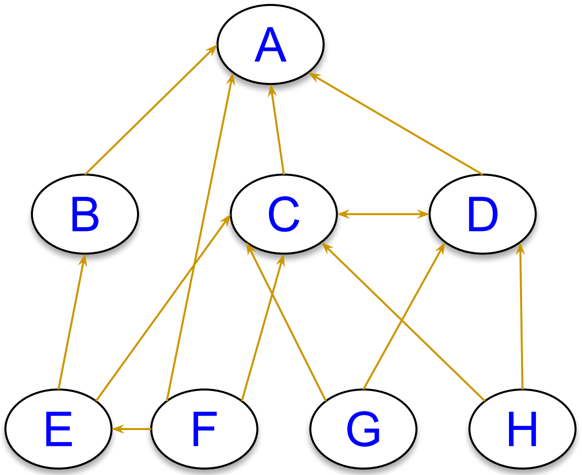
- Different influence trees possible for the same cascade
- A retweeting user may get exposed to the cascade through multiple friends

Problem statement

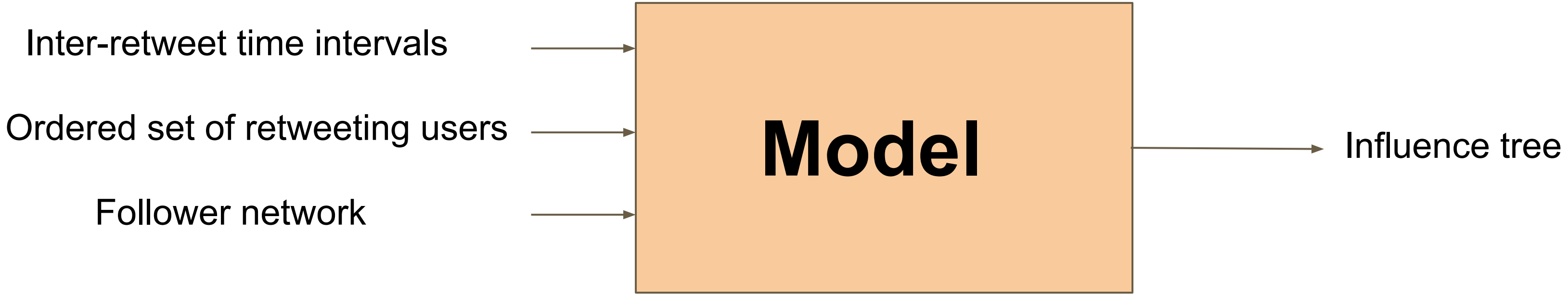
- Estimate the structure of influence tree for a cascade from temporal data
- Investigate the role of inter-retweet time intervals to construct influence tree



Time series



Follower network

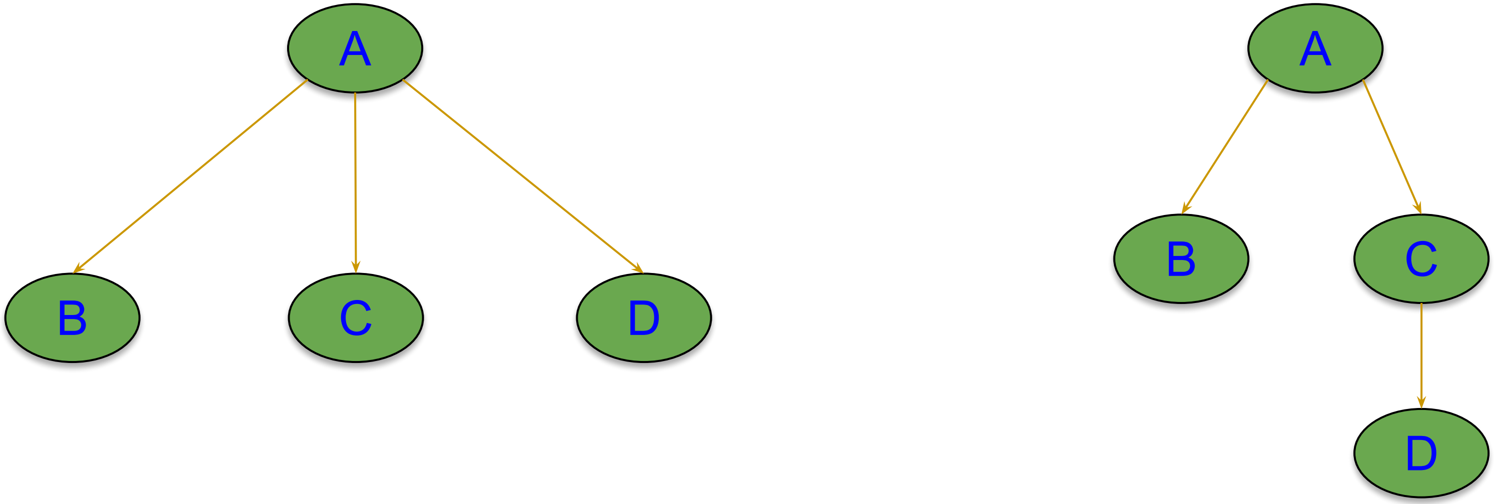


Outline

- **Characterization of influence trees**
- Formation of ground truth influence tree
- CasCon: Model to construct influence trees
- Analytical formulation of CasCon
- Experimental evaluation
- Conclusion

Characterization of influence trees

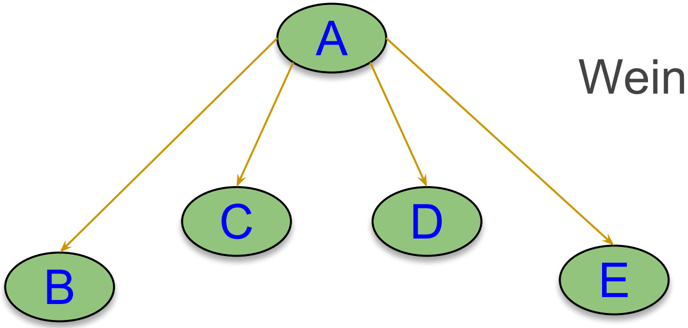
- How to distinguish between structures of 2 different influence trees?



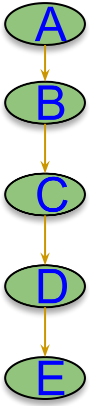
- Easy to distinguish for smaller cascades as shown
- Difficult for real time large cascades

Characterization of influence trees

- Define a set of structural metrics
  - Weiner index
  - Mean depth
  - Fraction of leaves
  - Average number of retweets
- Compute Weiner index for a tree of size n
  - Maximum (=1) for a line structure
  - Minimum (=0) for a star structure

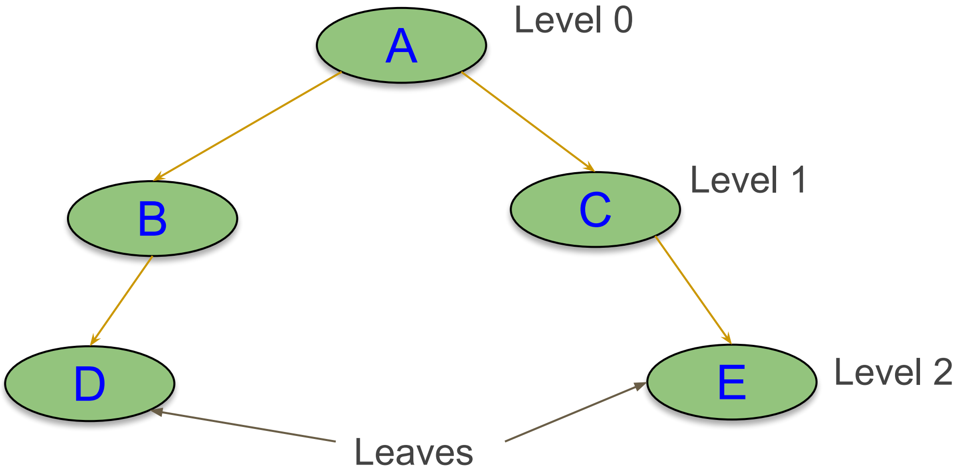


Weiner index = 0



Weiner index is 1

Characterization of influence trees



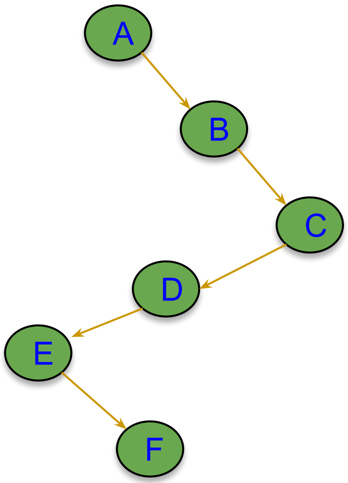
Mean Depth: (1+1+2+2)/4 = 1.5

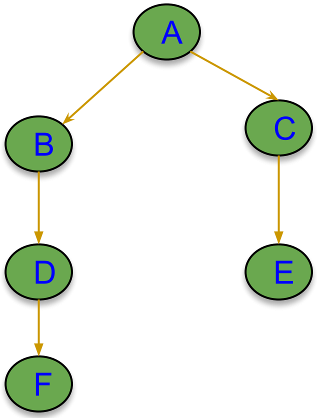
Average no. of retweets: (2+2)/2 = 2

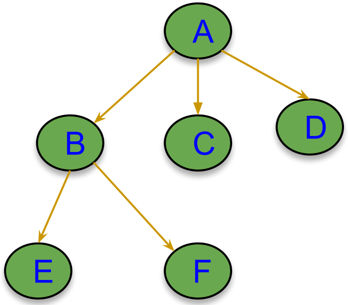
Fraction of leaves: 2/4 = 0.5

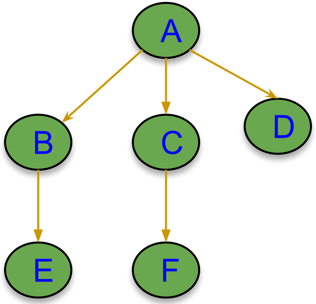
Characterization of influence trees

- Why do we need different structural metrics?









Weiner index = 1

Average no. of retweets = 2.5

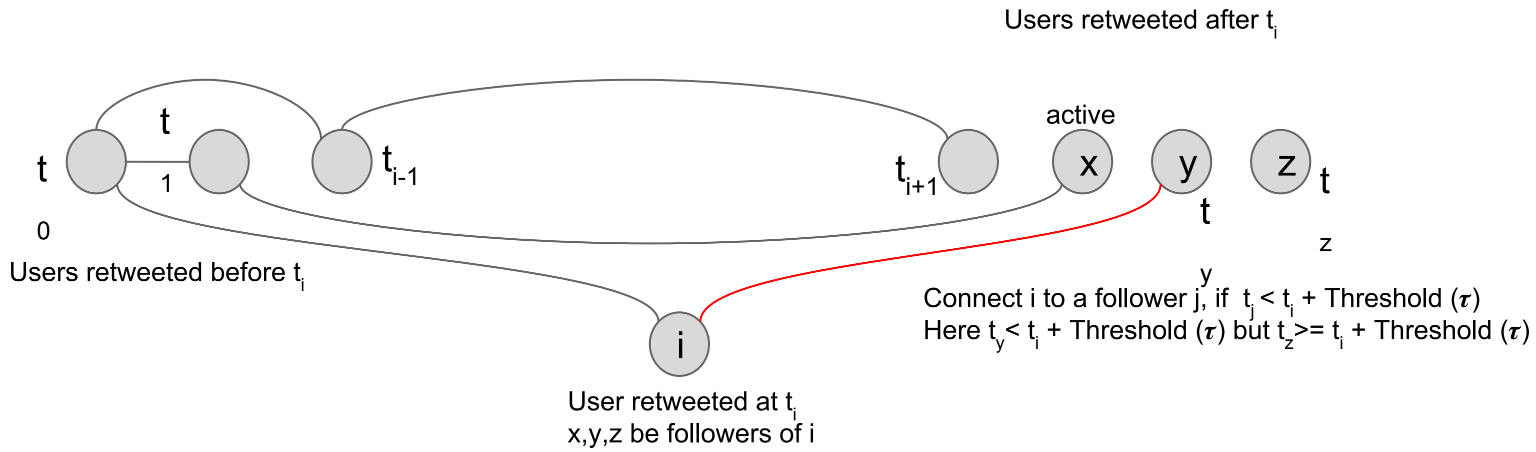
Mean depth = 1.4

Outline

- Characterization of influence trees
- **Formation of ground truth influence tree**
- CasCon: Model to construct influence trees
- Analytical formulation of CasCon
- Experimental evaluation
- Conclusion

Formation of ground truth influence tree

- Independent Cascade Model (ICM) [1]



- A retweeting user v^C influences an inactive follower u^C that retweets within a Threshold time limit (τ) after retweet by v^C

[1] D. Kempe, J. M. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network.,” pp. 105–147, ToC, 2015.

Formation of ground truth influence tree

Linear Threshold Model (LTM) [1]

- A retweeting user v^C is influenced by each of her active friend w^C who retweeted earlier in a cascade C according to an edge weight ~ *U[0,1]*
- A random threshold θ_{vC} ~ *U[0,1]* is chosen for v^C
- v^C retweets when sum of edge weights for its currently active friends exceeds θ_{vC}
- The most recent active friend is designated as the influencer of v^C

[1] K. AlFalahi, Y. Atif, and A. Abraham, “Models of influence in online social networks,” pp. 161–183, IJISA, 2014

Comparison between ground truth influence trees

- Comparison between gold standard influence trees with ICM & LTM based on Mean Absolute Error (MAE) and Mean Square Error (MSE)
- We establish compliance of only 46% between the ground truth influence trees where influencer-influencee node pairs exhibit consistency across both ICM & LTM models

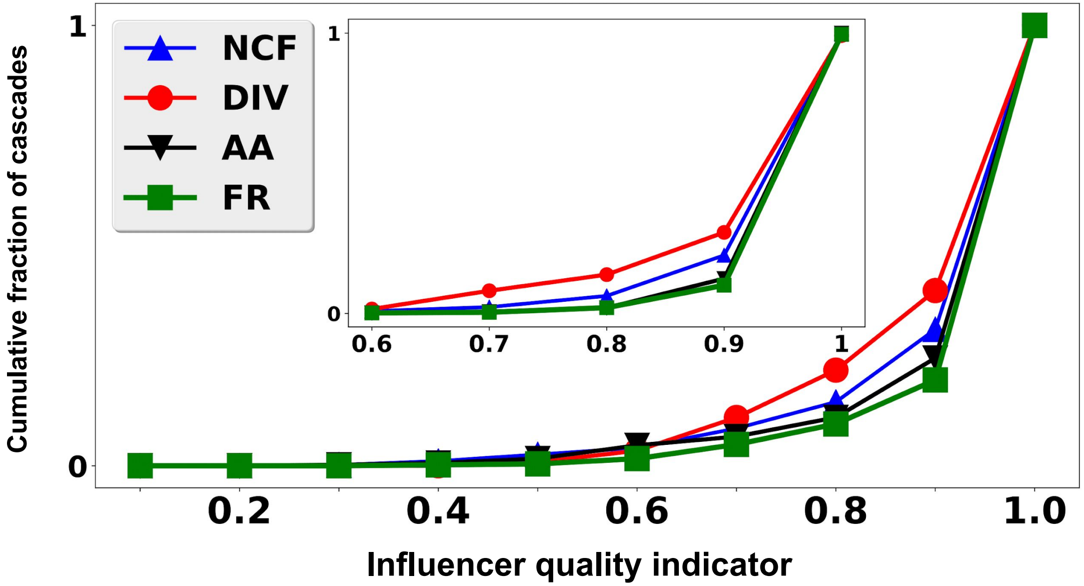
Model	**Wiener Index**	**Mean Depth**	**Fraction of Leaves**	**Average number of retweets**
MAE	0.21	0.10	0.12	0.23
MSE	0.11	0.06	0.04	0.14

Validation of ground truth influence tree

- Introduce a suite of influence metrics for pairwise influence between node pairs:
  - Normalized co-occurrence frequency (NCF) - No of cascades in which both influencer-influencee retweets
  - Diversity of co-occurrence (DIV) - Captures diversity in retweet times for cascades where both influencer-influencee participates
  - Adamic/Adar (AA) - Measures inverse log of follower count of each common friend between influencer-influencee pair
  - FollowerRank (FR) - Proportion of followers among all neighbors (friend+follower) of an influencer node

Validation of ground truth influence tree

- Influencer quality indicator I_Q - influence scores of ground truth influencer nodes
- 2 ways for validation of ground truth influence trees:
  - All influencer nodes across influence trees in the dataset
  - Validation of only the consistent influencer-influencee node pairs (46% of all pairs)



Observations:

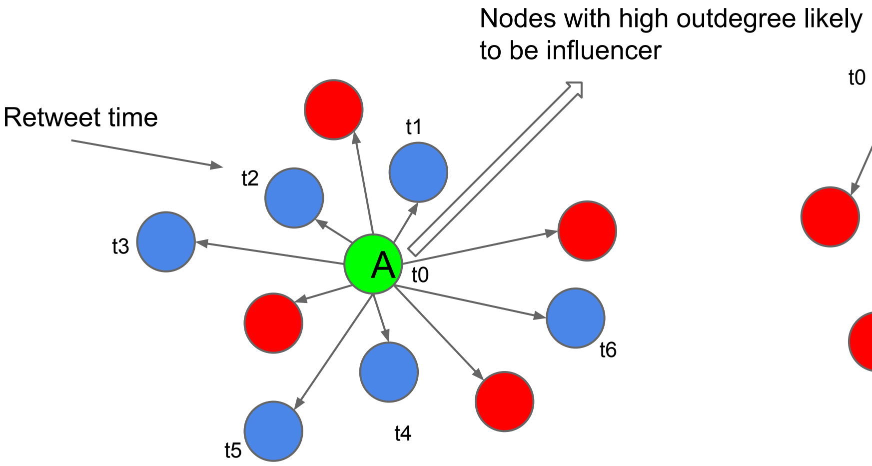
1) I_Q>=0.8 for 85% cascades in case of all influencer nodes
2) I_Q>=0.8 for 93% cascades in case of only consistent influencer-influencee node pairs

3) High influence scores of ground truth influencer nodes validate the influence trees obtained from ICM & LTM

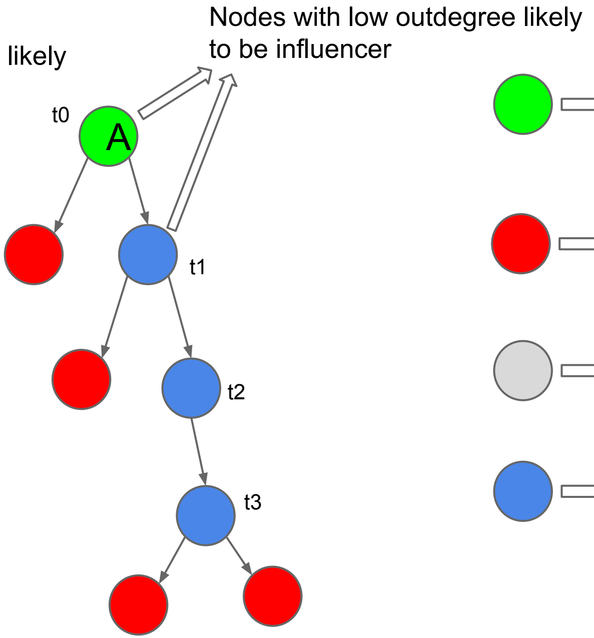
Outline

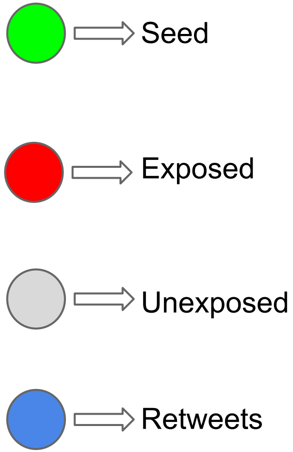
- Characterization of influence trees
- Formation of ground truth influence tree
- **CasCon: Model to construct influence trees**
- Analytical formulation of CasCon
- Experimental evaluation
- Conclusion

CasCon: Key idea



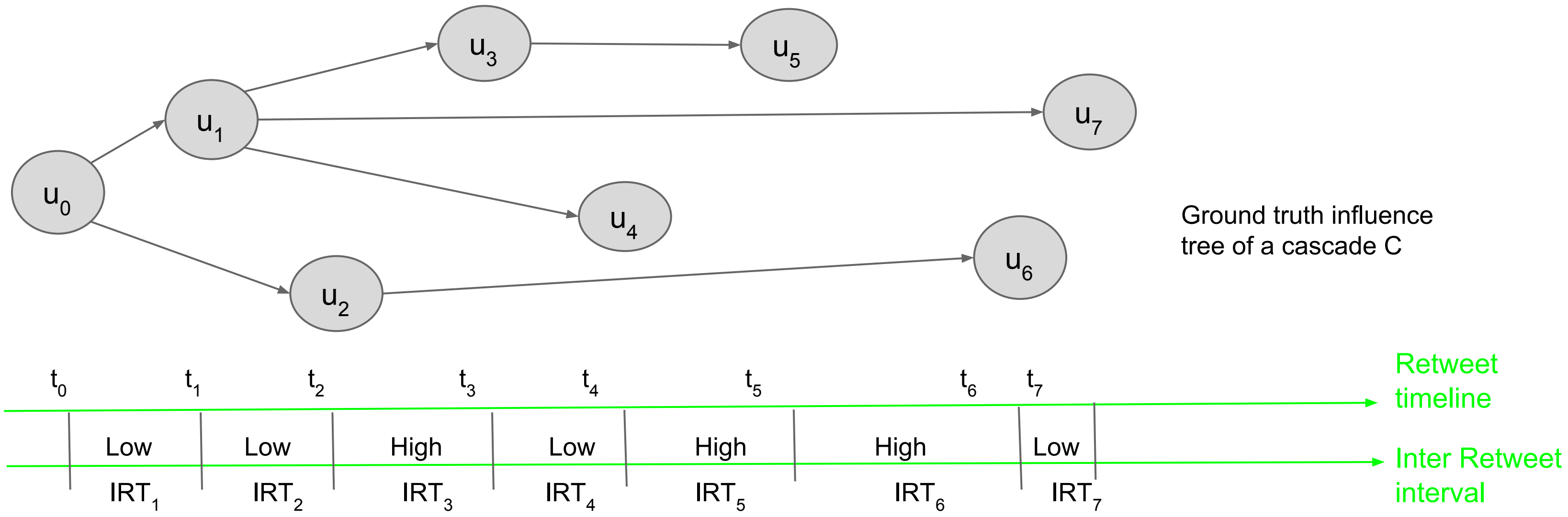
STAR NETWORK
successive retweets t_i - t_{i-1} can occur
with small inter-retweet time intervals





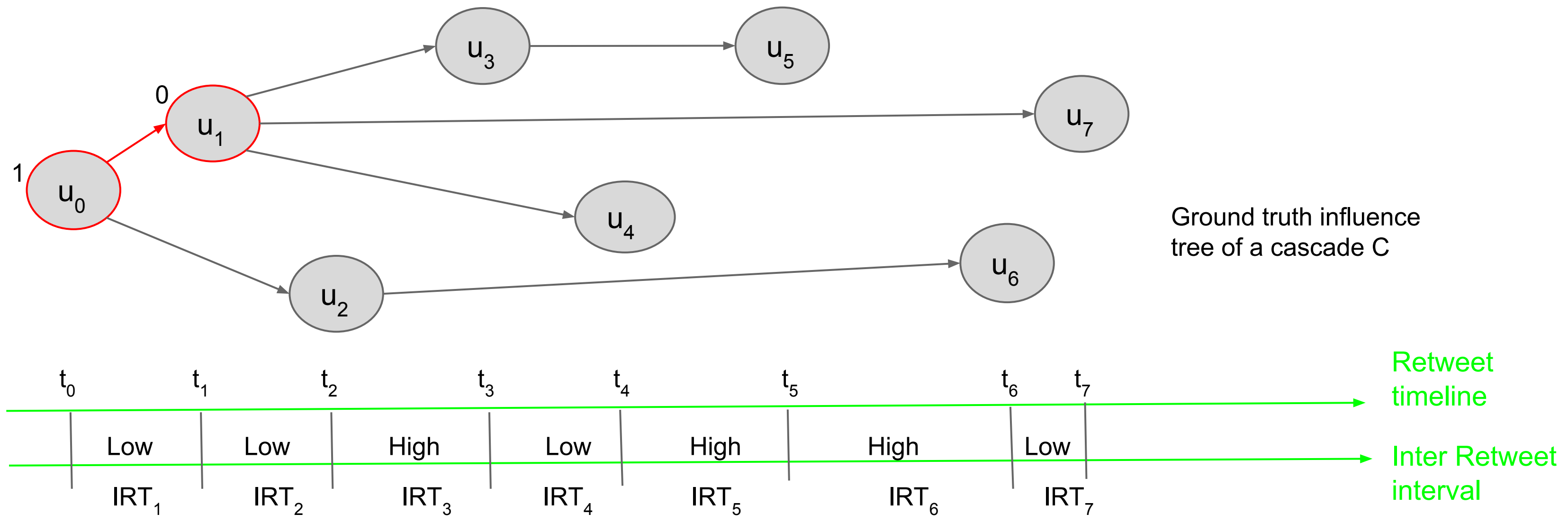
LINE NETWORK
successive retweets have high
inter-retweet time intervals

CasCon: Key idea



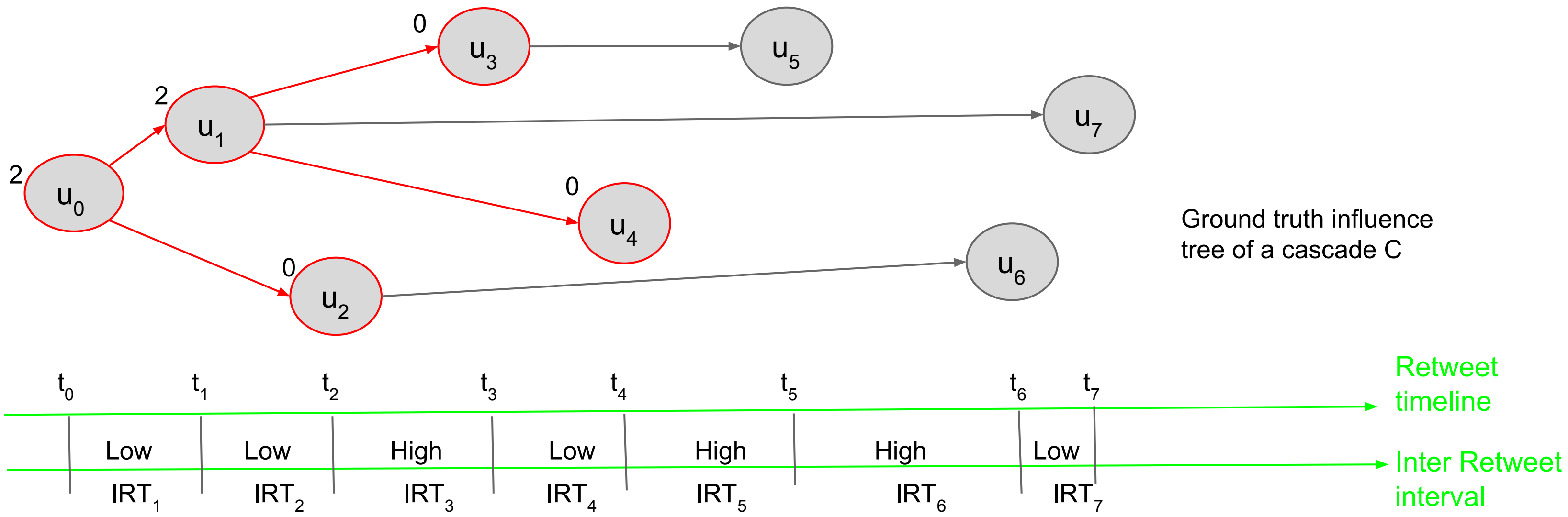
- Inter-retweet time intervals obtained from time series, classified as either high or low using simple outlier detection technique

CasCon: Key idea



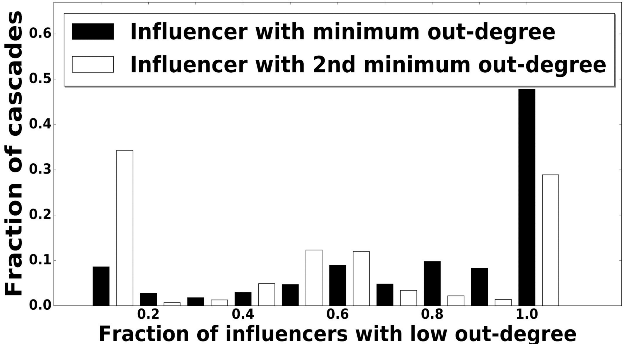
- At time t_2 , user u_2 enters the influence tree formed till then (marked by red)
- Possible influencers/ friends of u_2 = { u_0, u_1 } via follower network
- Since inter-retweet interval = Low at t_2, u_2 is connected to high outdegree node i.e. u_0 in the influence tree

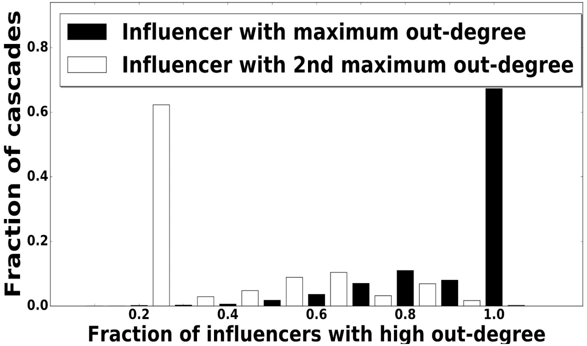
CasCon: Key idea



- At time t_5, user u_5 enters the influence tree formed till then (marked by red)
- Possible influencers/ friends of u_5 = { u_1, u_2, u_3 } via follower network
- Since inter-retweet interval = High at t_5, u_5 is connected to a low outdegree node i.e. u_3 in the influence tree

CasCon: Evidence from empirical data





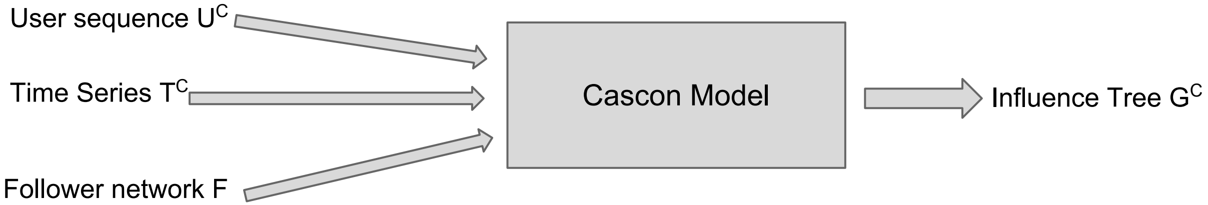
Observation: 86% of cascades have fraction of users with high IRT connecting to a node with low current out-degree >= 0.8

Observation: 77% of cascades have fraction of users with low IRT connecting to a node with high current out-degree >= 0.8

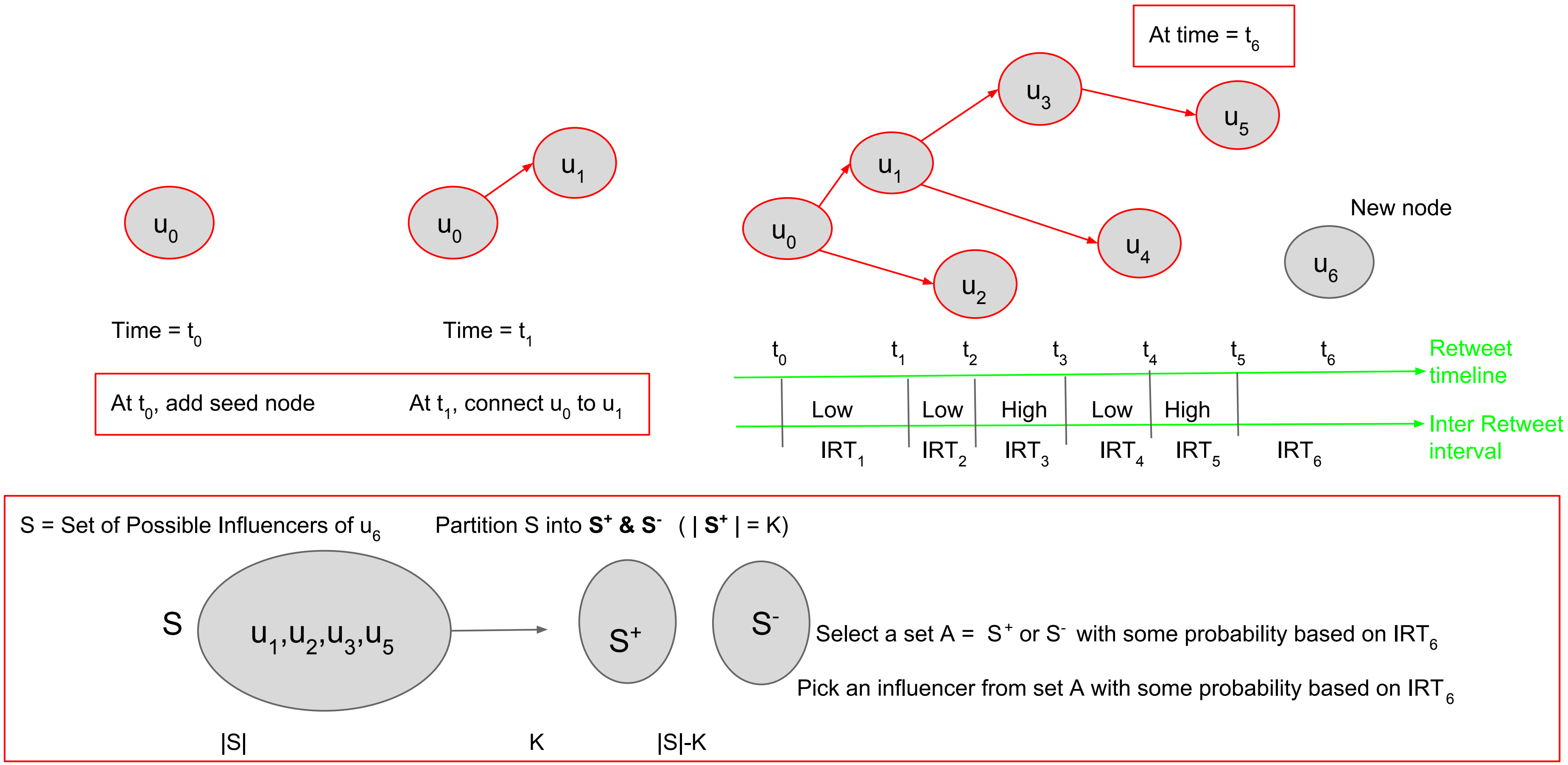
- **Key idea:** If the inter-retweet time interval is low, connect the newly retweeting user to the friend with high current out-degree and vice versa

CasCon: Model to construct influence trees

- Time series of cascade C T^C : $T_0^C, T_1^C, T_2^C, \ldots\ldots T_N^C$
- User series of cascade C U^C : $U_0^C, U_1^C, U_2^C, \ldots\ldots U^C_N$
- F : Follower information of U^C
- Select a model parameter K

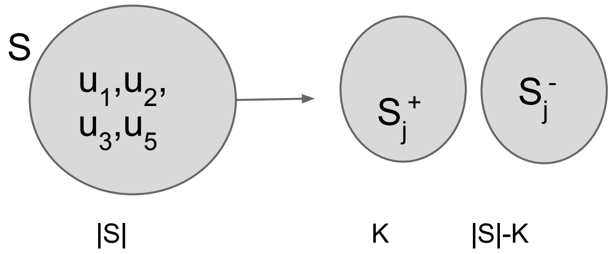


CasCon: Model to construct influence trees



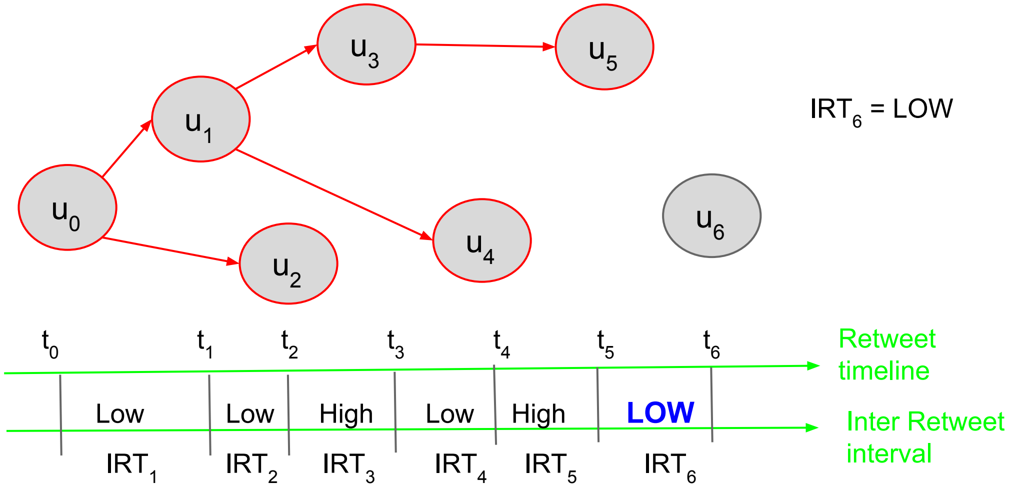
CasCon: Model to construct influence trees

S = Possible influencers of u_6



- S_j^+ **= Top-K users (based on outdegree) in S**
- out_{Sj+} = sum of out degrees of users in S_j^+
- out_{Sj-} = sum of out degrees of users in S_j^-

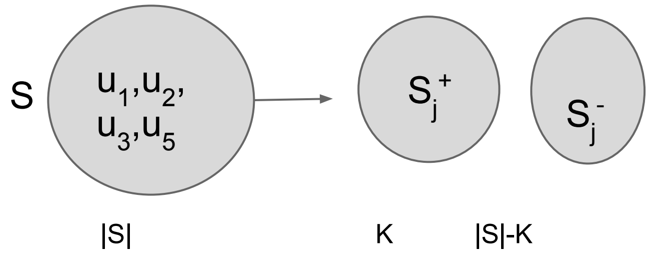
$$P(S_j^+) = K^*out_{Sj+} / (K^*out_{Sj+} + (|S|-K)^*out_{Sj-})$$



- Pick A = S_j^+ or S_j^-
- Pick user u_j in A with $P(u_j) \propto$ out-degree(u_j)
- Favours an influencer with high current out-degree

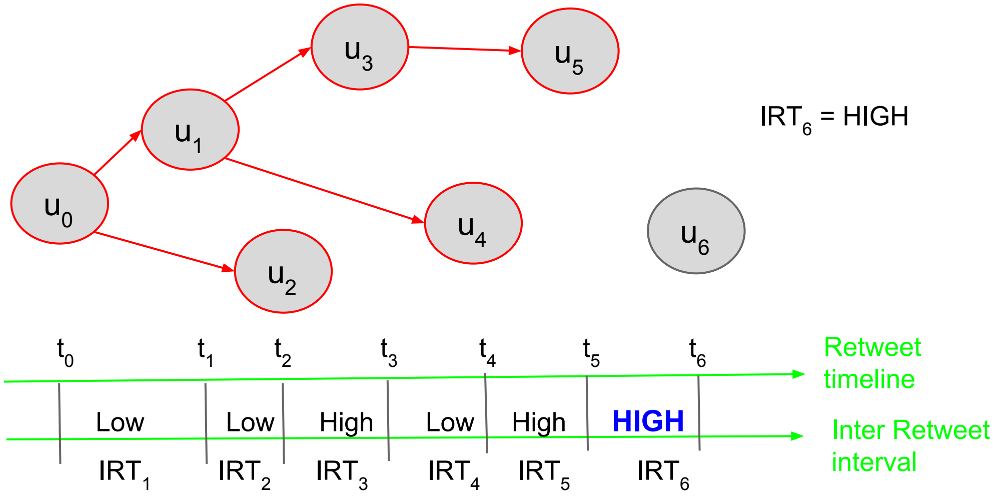
CasCon: Model to construct influence trees

S = Possible influencers of u_6



- **S_j^+ = Bottom-K users (based on outdegree) in S**

$$P(S_j^+) = K*(1/out_{Sj+}) / (K*(1/out_{Sj+}) + (|S|-K)*(1/out_{Sj-}))$$



- Pick A = S_j^+ or S_j^-
- Pick user u_j in A with $P(u_j) \propto 1/\text{out-degree}(u_j)$
- Favours an influencer with low current out-degree

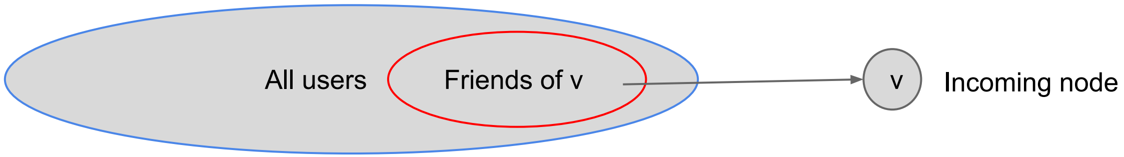
Outline

- Characterization of influence trees
- Formation of ground truth influence tree
- CasCon: Model to construct influence trees
- **Analytical formulation of CasCon**
- Experimental evaluation
- Conclusion

Analytical formulation of CasCon

- Derive degree distribution of predicted influence trees
- We rely on Barabasi-Albert preferential attachment model

Partial Cascade tree



- Case 1 : If IRT_v = Low, connect to a friend i (out-degree = k_i) of v with

$$\boxed{A(k_i) \;=\; (k_i+1) \,/\, \Sigma_{\text{friends of v}}\, k_j}$$

- Case 2 : If IRT_v = High, connect to friend i (out-degree = k_i) with

$$\boxed{A(k_i) \;=\; (\, 1/(k_i+1) \,) \,/\, \Sigma_{\text{friends of v}}\, 1/k_j}$$

Analytical formulation of CasCon

- Let Probability (IRT_v = Low) = q ; Probability (IRT_v = High) = 1-q

- Connect to a friend i (out-degree = k_i) of j with

$$\boxed{A(k_i) = q * (k_i+1) / \Sigma_{\text{friends of } v}\, k_j + (1-q) * (1/(k_i+1)) / \Sigma_{\text{friends of } v}\, 1/k_j}$$

- Rate of change of out degree of friend "i" =

$$\boxed{\frac{\partial k_i}{\partial t} = q\frac{k_i+1}{\sum_j k_j} + (1-q)\frac{\frac{1}{k_i+1}}{\sum_j \frac{1}{k_j}}}$$

$k_i(T_i) = 0$

$$\boxed{k_i(t) = \frac{4}{3q-5}\left[t\left(\frac{t}{t_i}\right)^{\frac{3q-5}{4}} - \frac{1}{2}(1-q)t\right]} \xrightarrow{p(k) = \frac{\partial P(k_i(t)<k)}{\partial k}} \boxed{p(k) = \frac{1}{t+1}\left[\frac{(3q-5)k}{4t} + \frac{1-q}{2}\right]^{\frac{9-3q}{3q-5}}}$$

Analytical formulation of CasCon

- Fraction of leaves : **FL = P(k_{min}) for n*P(k_{min}) > 0** n = total no.of nodes in the network
- Wiener Index : **WI = (1 + O(1)) (log n/ log ~d)** where $$\tilde{d} = \frac{\sum_{k_{min}}^{k_{max}} k^2 nP(k)}{\sum_{k_{min}}^{k_{max}} knP(k)}$$ [1]
- Approximate height : **h = $\lceil \log_{kmax} ((k_{max} - 1)*n + 1) - 1 \rceil$**
- Mean Depth : $$M^C = \frac{hz^{h+2} - z^{h+1}(h+1) + z}{n(z-1)^2}$$ $z = \Sigma_{k = kmin}^{kmax} \; k*P(k)$ [2]
- Average Retweets : **AR = n/h** h = approximate height of the tree as above

[1] F. Chung and L. Lu, “The average distance in a random graph with given expected degrees,” Internet Mathematics, pp. 91–113, 2004.
[2] W. contributors, “K-ary tree — wikipedia, the free encyclopedia,” 2018.

Outline

- Characterization of influence trees
- Formation of ground truth influence tree
- CasCon: Model to construct influence trees
- Analytical formulation of CasCon
- **Experimental evaluation**
- Conclusion

Baseline algorithms

- **StrucVir:** Find influencer based on official retweet & accredited repost heuristics [1]

- **InfLoc:** Employ a Random Walk With Restart approach to compute pairwise influence [2]

- **Cascon-Base:** Same as Cascon but instead of dividing into S^+, S^- sets , the influencer is directly picked with probability based on high/low inter retweet time interval

[1] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, "The structural virality of online diffusion," pp. 180–196, Manag. Sci, 2015.
[2] J.Zhang, J.Tang, J.Li, Y.Liu and C.Xing, "Who influenced you? predicting retweet via social influence locality," TKDD, 2015..

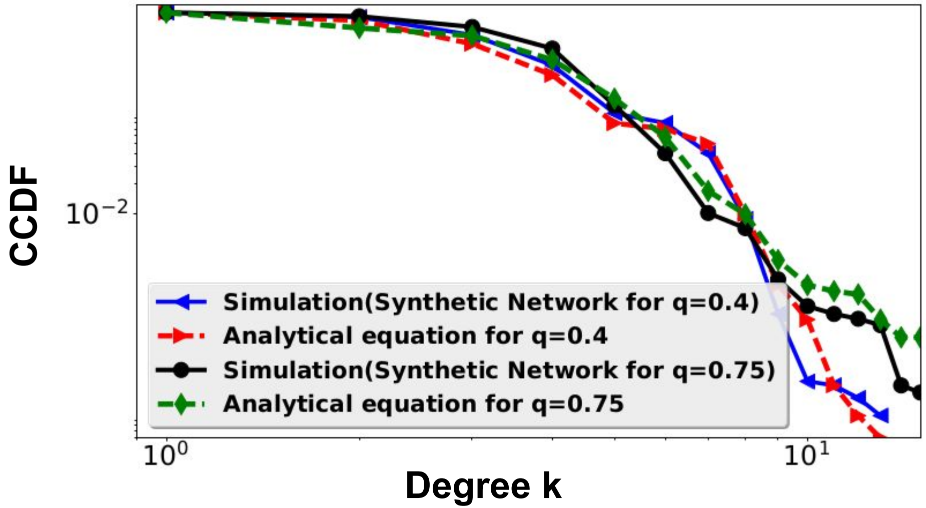
Macroscopic evaluation of CasCon

Evaluation of analytical formulation on synthetic network

- Cascade trees simulated over synthetic network of 10000 nodes using attachment probability :

probability(new node connects to friend i) = $q * (k_i+1) / \Sigma_{\text{friends of } v}\, k_j \quad + \quad (1-q) * (1/(k_i+1)) / \Sigma_{\text{friends of } v}\, 1/k_j$

- Comparing the complementary cumulative degree distribution calculated empirically with analytic P(k)



Observation: Degree distribution of simulated influence trees over synthetic network closely resembles analytically derived p(k)

Macroscopic evaluation of CasCon

Evaluation of analytical formulation on synthetic network

- Structural metrics computed empirically for the cascades generated from synthetic network ----[1]
- Structural metrics computed from analytic derivation using P(k) for the above trees ----[2]
- Mean Absolute error of metrics computed from [1] and [2] for different values of q for analytic model

Probability(q)	Wiener Index (MAE)	Mean Depth (MAE)	Fraction of leaves (MAE)	Average no. of retweets (MAE)
0.01	0.373	0.660	0.663	0.071
0.40	0.384	0.390	0.530	0.077
0.75	0.664	0.410	0.420	0.08

Observation:

1) High agreement between simulation and analytical method
2) Exceptionally high similarity for metrics such as Weiner index and Average no. of retweets

Macroscopic evaluation of CasCon

Evaluation of analytical formulation on real cascades

Cascade size	Wiener Index (MAE)	Mean Depth (MAE)	Fraction of leaves (MAE)	Average Retweets (MAE)
All	0.30	0.080	0.170	0.05
>100	0.21	0.064	0.060	0.02

Comparison with LTM

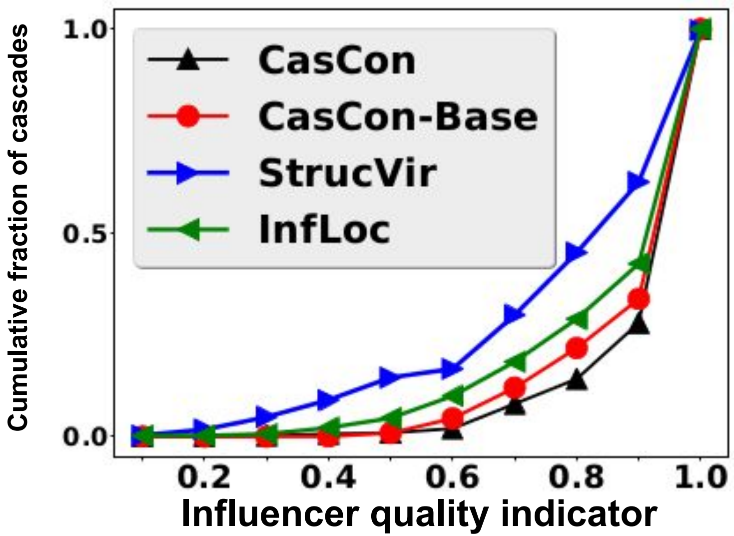
Cascade size	Wiener Index (MAE)	Mean Depth (MAE)	Fraction of leaves (MAE)	Average Retweets (MAE)
All	0.30	0.080	0.170	0.05
>100	0.21	0.064	0.060	0.02

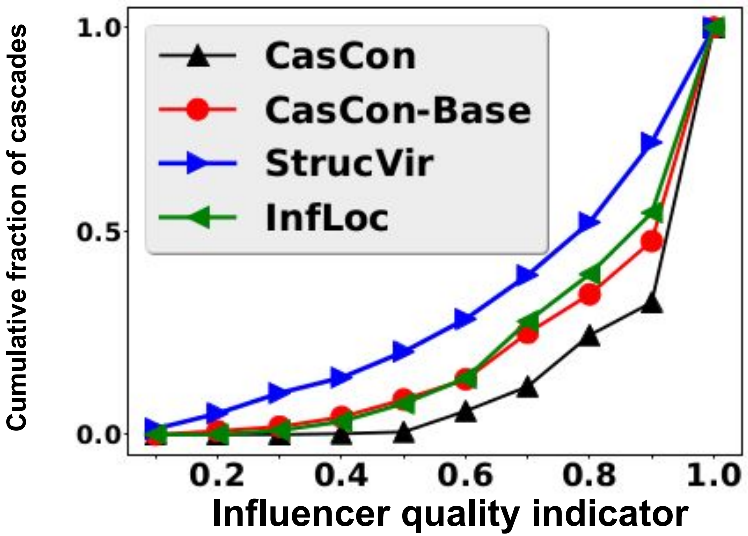
Comparison with ICM

Observation: Analytical formulation closely approximates ground truth influence trees for large cascades due to presence of more retweets

Microscopic evaluation of CasCon

Evaluation based on influence measures





Observation: 1) 75% of cascades have I_Q >= 0.8 for the *Normalized co-occurrence frequency* measure

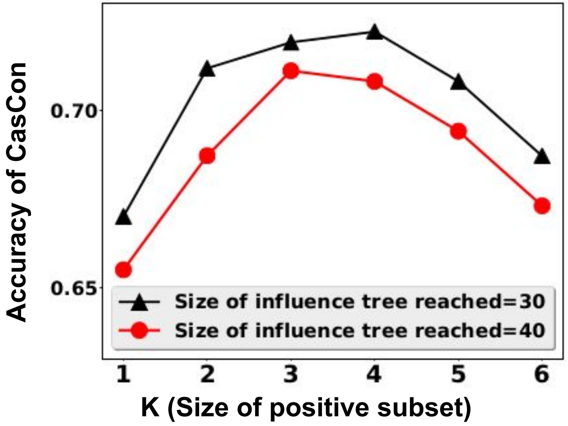
2) Only 48% and 60% of cascades have I_Q >= 0.8 for *StrucVir* and *InfLoc*

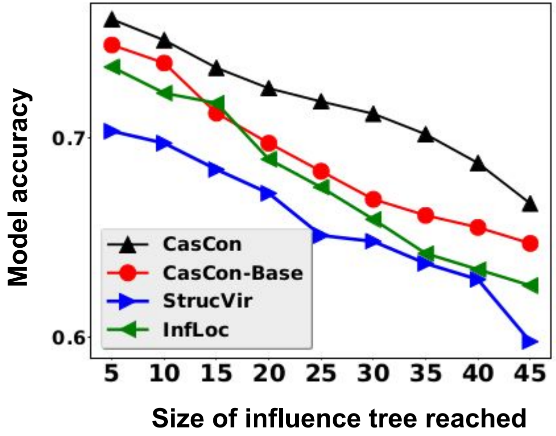
Observation: 1) 85% of cascades have I_Q >= 0.8 for the *Adamic/Adar* measure

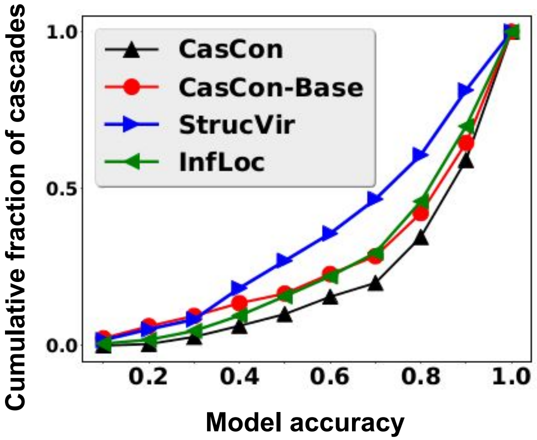
2) Only 54% and 71% of cascades have I_Q >= 0.8 for *StrucVir* and *InfLoc*

Microscopic evaluation of CasCon

Evaluation based on model parameters







Observation: 1) Accuracy of selecting correct influencer increases with size of positive subset K for *CasCon*

2) Increasing K beyond a limit reduces performance

Observation: 1) *CasCon* outperforms baselines in selecting correct influencer over different sizes of influence tree

2) Accuracy drops with increase in cascade size

Observation: 1) Model accuracy >=0.8 for 73% of cascades in case of *CasCon*

2) Model accuracy >=0.8 for only 32% and 60% cascades in case of *StrucVir* and *InfLoc*

Outline

- Characterization of influence trees
- Formation of ground truth influence tree
- CasCon: Model to construct influence trees
- Analytical formulation of CasCon
- Experimental evaluation
- **Conclusion**

Conclusion

- Obtained ground truth structure of influence trees from standard diffusion models
  - Validated the ground truth from a suite of influence metrics
- Defined various structural indices to quantify diverse influence trees
- Developed a simulation model to select the most likely influencer for each retweeting user
  - Predicts the influence tree of a cascade closely resembling ground truth with good accuracy
- Provide an analytical bound on the probability for a retweeting user to connect to its most likely influencer
  - Estimate the degree distribution of predicted influence trees
- Both simulation model and analytical formulation have been shown to correctly reconstruct ground truth influence trees with high accuracy
  - Outperforms competing algorithms

THANK YOU