Analyzing and Modeling Diffusion using Social Topic Inference Relationship Prediction

KOMODO

Mentor: Carl Yang

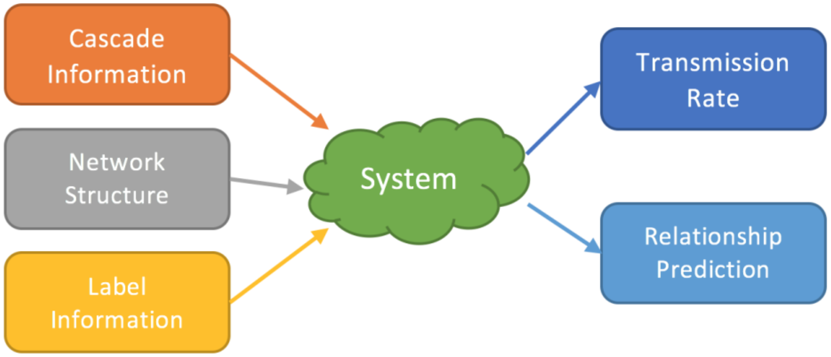
Problems & Goals

- **Information Diffusion** is understanding how or why information spread within the network (e.g. activation probability, transmission rate between nodes).

  In this project, we want to infer transmission rate between nodes to have better understanding on how fast information propagate within network.

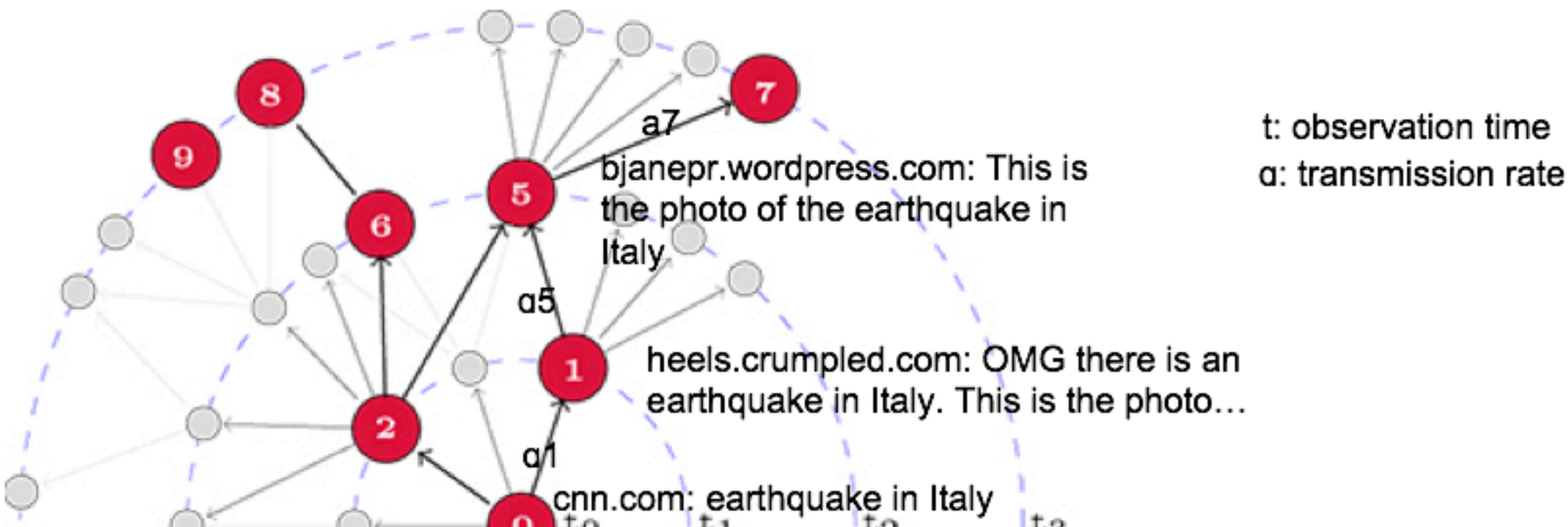
- And the **Relationships** will impact on how the information propagated.

  In this project we want to build a model for predicting relationship within nodes and use the information to improve the transmission rate prediction model.



Motivation

- Learning information diffusion is important to understand how a piece of information flows
- With the Topic Inference in combination with information diffusion model, we can specify the similarity of topic interests within the observation nodes
- This similarity can be used for making a recommendation network based on specific topic



Information Diffusion Model (NETRATE)

$$L(\mathbf{t}^1..\mathbf{t}^c;\mathbf{A}) = \sum \Psi_1(\mathbf{t}^c;\mathbf{A}) + \Psi_2(\mathbf{t}^c;\mathbf{A}) + \Psi_3(\mathbf{t}^c) \qquad (1)$$

where for each cascade $\mathbf{t}^c \in \{t^1, ..., t^c\}$, each function can be derived into

$$\Psi_1(\mathbf{t}^c;\mathbf{A}) = \sum_{i:t_i \leq T} \sum_{t_m > T} \log S(T|t_i;\alpha_{i,m})$$

$$\Psi_2(\mathbf{t}^c;\mathbf{A}) = \sum_{i:t_i \leq T} \sum_{t_j < t_i} \log S(t_i|t_j;\alpha_{i,m})$$

$$\Psi_3(\mathbf{t}^c;\mathbf{A}) = \sum_{i:t_i \leq T} \log \sum_{j:t_j < t_i} H(t_i|t_j;a_{j,i})$$

Transmission Likelihood. For this experiment we use Exponential Model transmission likelihood f given as

$$f(t_i|t_j;\alpha_{j,i}) \begin{cases} \alpha_{j,i}.\exp^{-\alpha_{j,i}(t_i - t_j)} & \text{if } t_j < t_i \\ 0 & \text{otherwise} \end{cases}$$

Survival function. S is a probability of a node survives uninfected until time T. Given the transmission likelihood, we can derive the survival function of our equation as

$$\log S(t_i|t_j;\alpha_{j,i}) = -\alpha_{j,i}(t_i - t_j)$$

Hazard function. H is a Hazard function or instantaneous infection rate of edge $j \rightarrow i$. Given the exponential transmission likelihood model, we get our hazard function as

$$H(t_i|t_j;\alpha_{j,i}) = \alpha_{j,i}$$

Limitation: This algorithm just consider the time and the cascades itself without other latent variables

Solution: NETRATE + Topic Inference

Expectation Maximization (EM). To infer the topics we proposed Expectation Maximization (EM) multinomial topic modeling [1] to get expectation probability of words p_j, k by using calculation

$$\varrho(\theta;\theta^{(n)}) = \sum_{ij}(\sum_{k} x_{i,k} \log p_{j,k}) + \log \pi_j$$

And then we minimize the variational free energy in the M-step using

$$p_j^{(n+1)} = \frac{\sum_i x_i w_{ij}}{\sum_i x_i^T 1 w_{ij}}$$

and

$$\pi_j^{(n+1)} = \frac{\sum_i w_{ij}}{N}$$

Topic 1 (war)	Topic 2 (motivation)	Topic 3 (peace)	Topic 4 (politic)	Topic 5 (opression)
mars	know	commitment	bein	january
jihad	efficient	applaud	fails	jiving
lose	divisi	king	political	signs
bureaucratic	looking	fulfill	diaz	silencing
para	karni	peace	kala	water
world	scourge	laptop	families	peace
Topic 6 (housing)	**Topic 7 (journal)**	**Topic 8 (history)**	**Topic 9 (game)**	**Topic 10 (army)**
using	journal	history	game	mars
standards	involves	values	bureaucratic	para
british	stairway	para	heartiest	conveying
tv	drilling	mars	fulfill	history
electricity	establishment	applaud	tap	world
history	precipitate	peace	prophet	applaud
Topic 11 (teamwork)	**Topic 12 (election)**	**Topic 13 (finance)**	**Topic 14 (danger)**	**Topic 15 (innovation)**
tarnishes	great	pale	danger	phones
applaud	hours	unsought	wait	rates
team	world	middle	mars	great
great	king	charge	country	imagine
good	tarnishes	neo	duties	january
federal	task	baby	federal	defense

$$f(t_i|t_j;\alpha_{j,i};\theta_{ck}) \begin{cases} \theta_{ck}.\alpha_{j,i}.\exp^{-\alpha_{j,i}(t_i-t_j)} & \text{if } t_j < t_i \\ 0 & \text{otherwise} \end{cases}$$

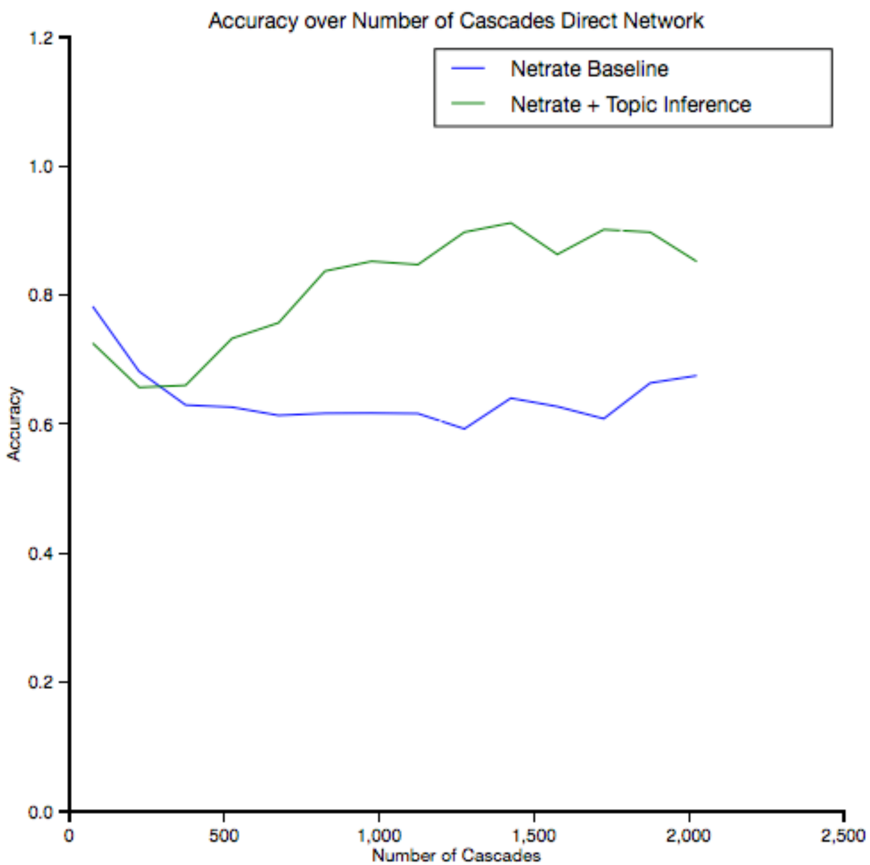
and the hazard function will be

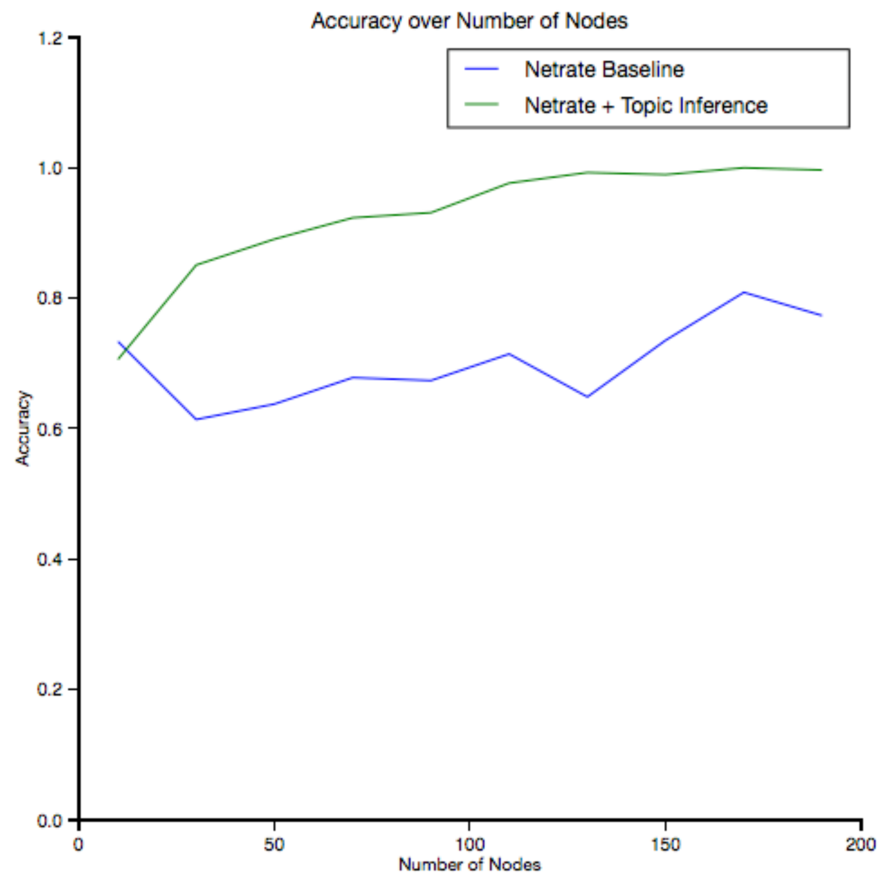
$$H(t_i|t_j;\alpha_{j,i};\theta_{ck}) = \alpha_{j,i}.\theta_{ck}$$

where θ_{ck} is a probability of the cascade c belong to the topic k. With this formula we can produce k number of $\alpha_{j,i}$ that will infer how fast the transmission rate between node $j \rightarrow i$ for that particular topics.

Experiments and Results

Testing the proposed solution for direct network





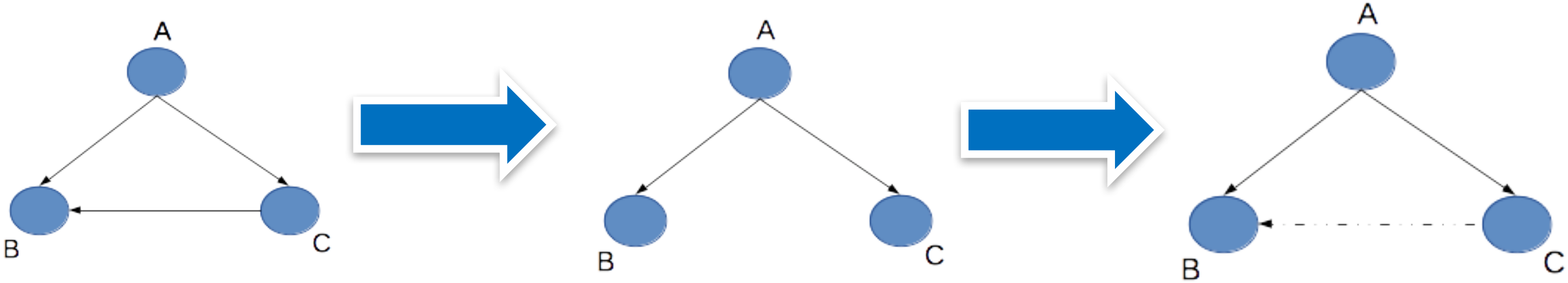
Experiments and Results

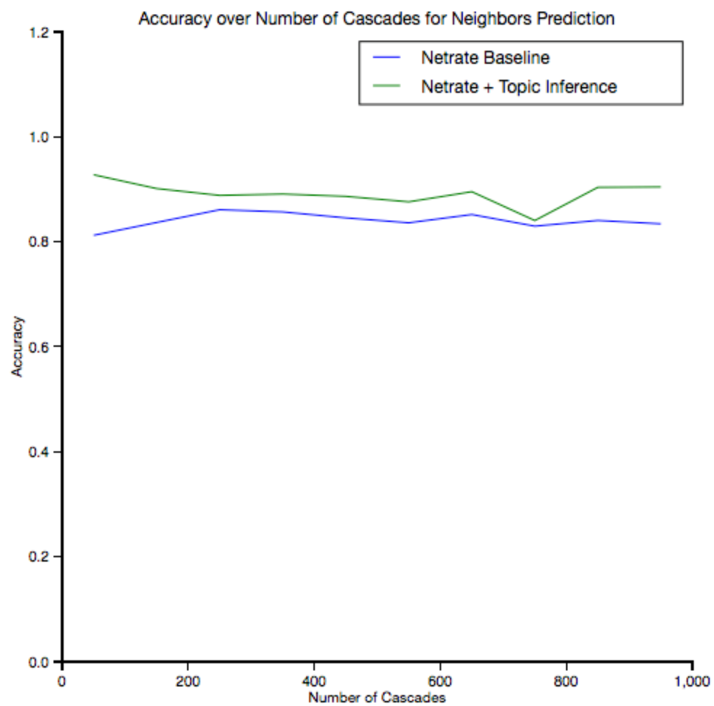
Testing our solution to predict neighbor network

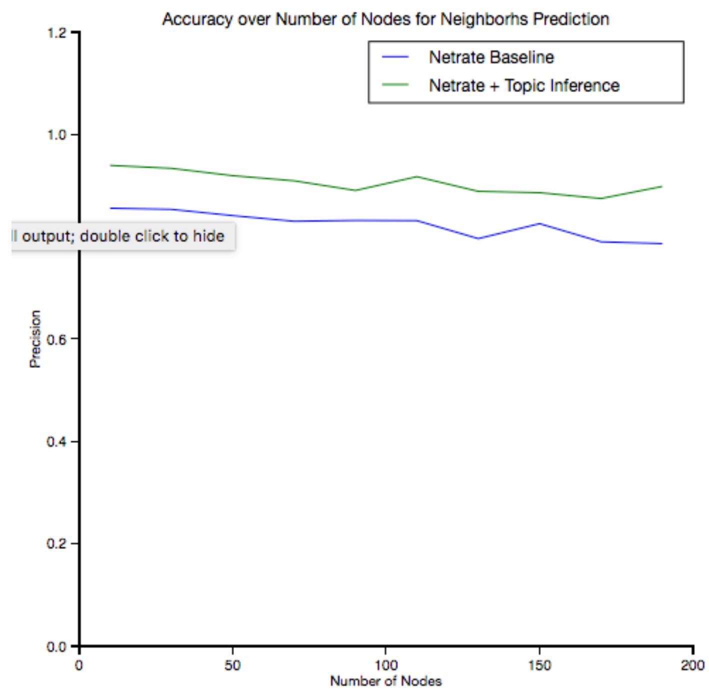
Suppose we have this graph

We hide the cascades from C

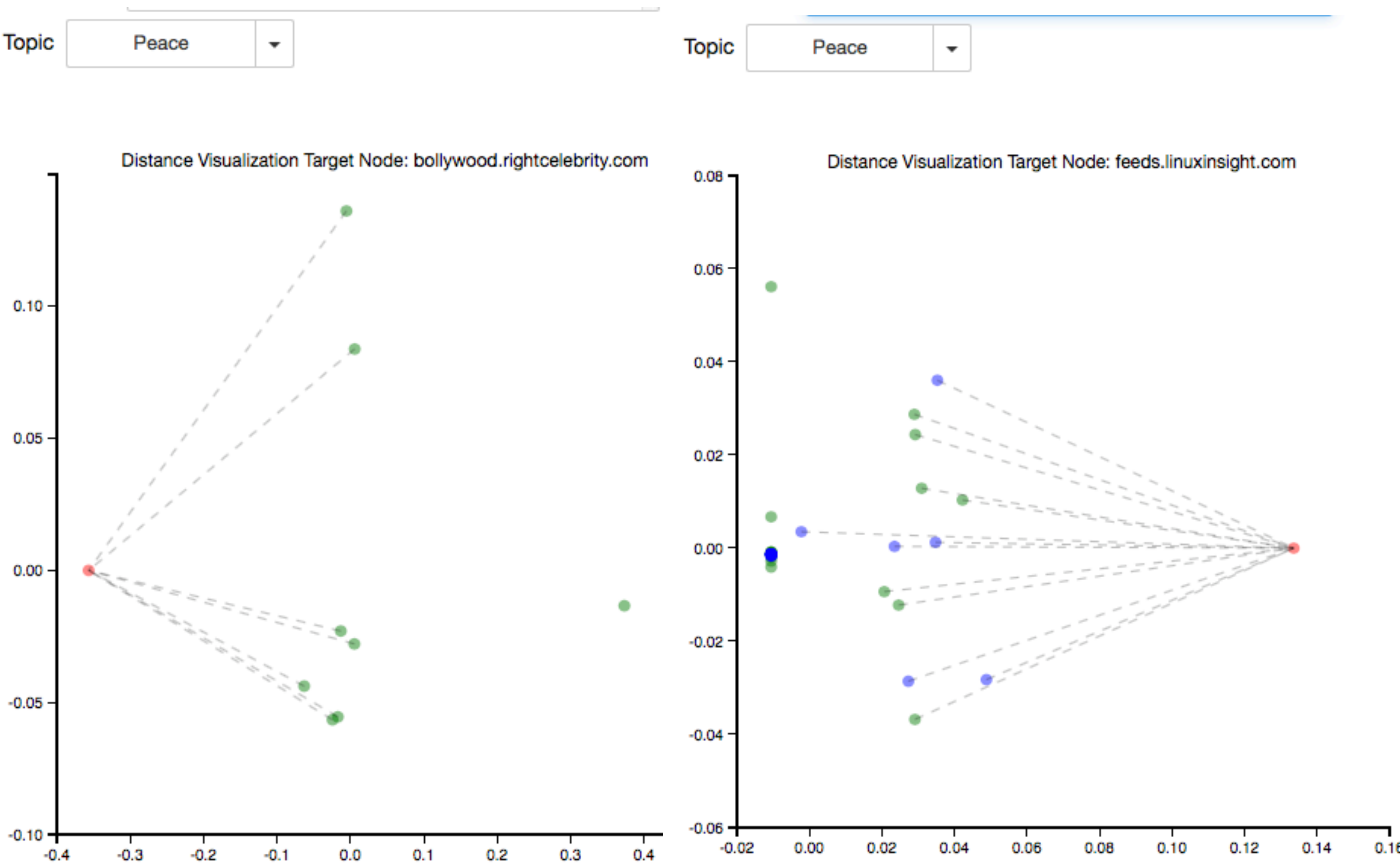
predict the edge within B and C
using the cascades from A







Demo


Topic
Peace
Distance Visualization Target Node: bollywood.rightcelebrity.com
0.10
0.05
0.00
-0.05
-0.10
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
Topic
Peace
Distance Visualization Target Node: feeds.linuxinsight.com
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.02
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

Future Work

- Implement the algorithm into real time network analysis using streaming to produce dynamic recommendation network
- Improve the topic modeling with supervised learning to give more precise prediction / label
- Improve the transmission rate prediction model with another parameters like user profile or geospatial location
- Combine the algorithm with graph database to visualize network structure using transmission rate prediction