Social Influence Locality for Modeling Retweeting Behaviors

Jing Zhang, Biao Liu, Jie Tang, Ting Chen and Juanzi Li
Department of Computer Science, Tsinghua Univeristy

Social Influence occurs when one's emotions, opinions, or behaviors are affected by others.

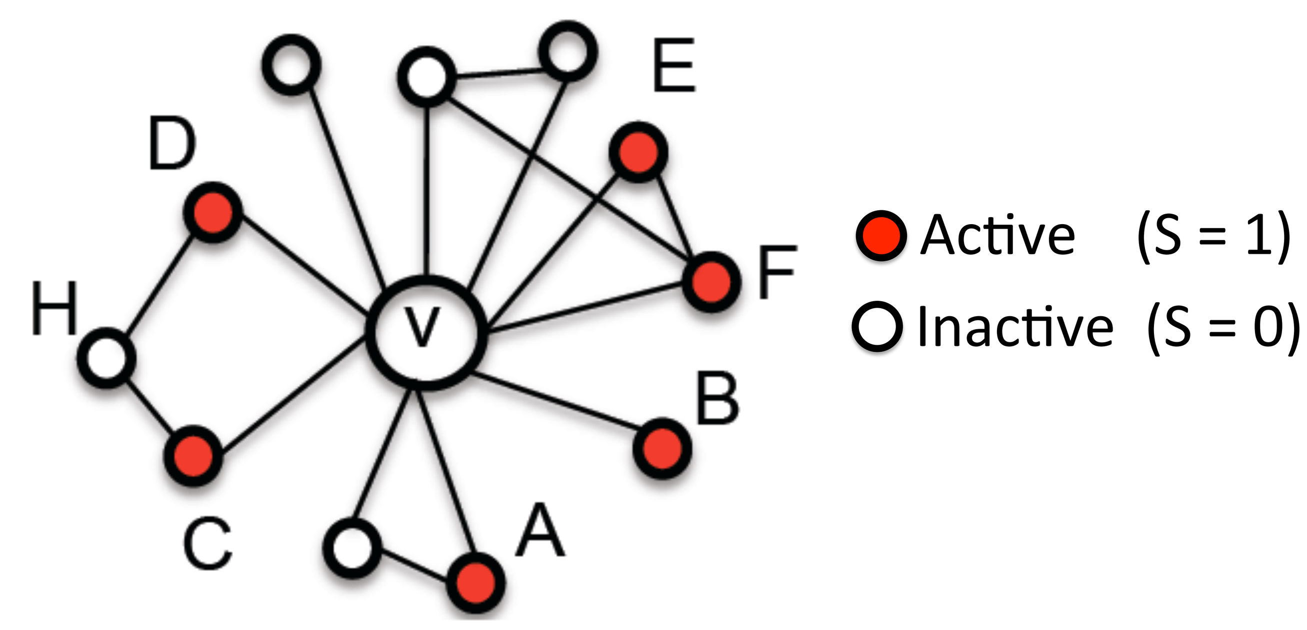
Influence is local in most cases such as retweet behavior.

Influence locality function:

$$Q(S_v, G_v^\tau), \quad with\ \tau \in \mathbb{N}^+$$

where G_v^τ is v's τ-ego network. S_v is the active neighbors in G_v^τ.

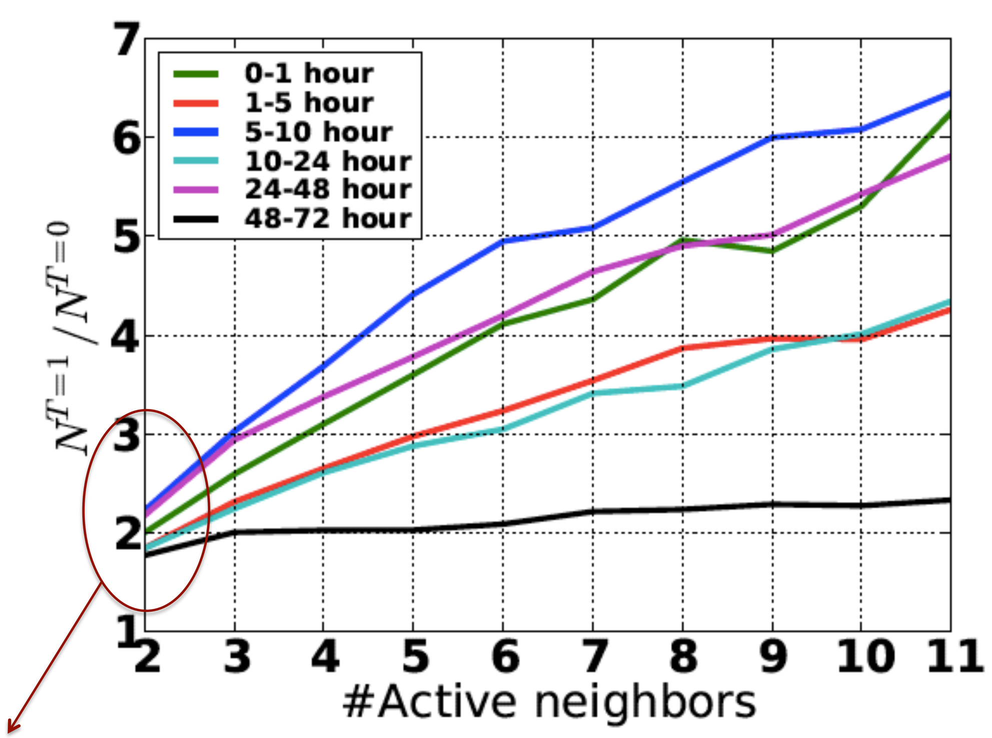
Part of v's 2-ego network



Influence Test

- Sina weibo Retweet Data: 1,776, 950 users, 308,489,739 follow relationships, 300,000 original microblogs, and 23,755,810 retweets.
- Test:
  - Treatment group: Users with >=1 active neighbors.
  - Control group: Users with =1 active neighbors.
- To avoid the selection bias:
  - For each user in treatment group, find the most matched user from the original control group.
  - Learn a logistic regression model to estimate the probability of each user being treated. Matching is based on the probability.
- To avoid the confounding bias:
  - Construct the two groups for each microblog and each time span after the microblog being published independently.

Test Results:



The fraction of active users with 2 active neighbors is about 2 times the fraction of active users with only 1 active neighbors.

Influence Measure

$$Q(S_v, G_v^\tau) = w \times g(S_v, G_v^\tau) + (1-w) \times f(S_v, G_v^\tau)$$

Pairwise Influence

F > B for influence on v ?
B only has one path to reach v
F has a number of paths to connect v

$$g(S_v, G_v^\tau) = \sum_{v_i \in S_v} p_{v_i}$$

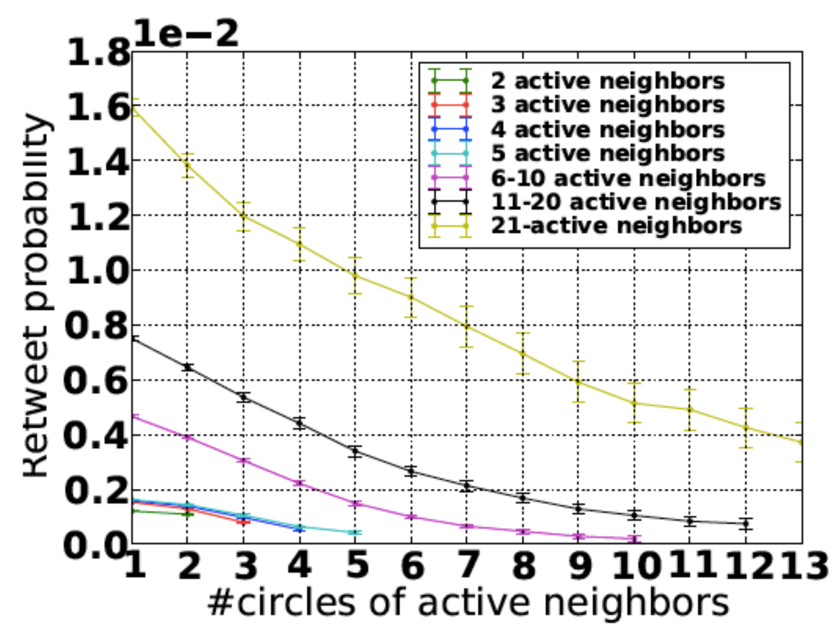
P_{vi} is random walk probability from v_i to v

Structure Influence

C+D > A+ B for influence on v ?
A and B distribute into different circles
C and D reside in the same circle

$$f(S_v, G_v^\tau) = e^{-\mu |C(S_v)|}$$

$C(S_v)$ is the number of circles formed by S_v



The retweet probability is negatively correlated with the number of circles.

Results

Performance of retweet behavior prediction (%)

Model	Prec.	Rec.	F1	Acc.
LRC-B	68.11	74.26	71.05	69.74
LRC-Q	66.82	77.22	71.65	69.44
LRC-BQ	69.89	77.06	73.30	71.93

LRC-B: Logistic regression classifier with only basic features (e.g., gender, verification status and so on).

LRC-Q: Logistic regression classifier with only influence locality function. (w=0.5, g=g_6)

LRC-BQ: Combine basic features and influence locality function together.

With only influence locality influence, we can obtain a F1-score of 71.65%.

Performance with and without structure influence (%)

Model	Prec.	Rec.	F1	Acc.
LRC-Q(w=1)	49.51	51.53	50.50	49.49
LRC-Q(w=0.5)	51.86	67.70	58.73	52.43

80% instances only have one active neighbors, thus the effect of structure influence can not be presented. So we sample instances with the number of active neighbors larger than 5.

Performance with different pairwise functions (%)

Model	Prec.	Rec.	F1	Acc.
$g_1 = \sum p_{v_i}$	57.42	77.13	65.83	59.96
$g_2 = \frac{\sum p_{v_i}}{\lvert S_v \rvert}$	60.21	75.03	66.81	62.72
$g_3 = \sqrt[\lvert S_v \rvert]{\prod p_{v_i}}$	60.28	75.31	66.96	62.84
$g_4 = \sum h_{v_i} p_{v_i}$	58.85	92.68	71.99	63.94
$g_5 = \frac{\sum h_{v_i} p_{v_i}}{\lvert S_v \rvert}$	61.57	91.72	73.68	67.24
$g_6 = \sqrt[\lvert S_v \rvert]{\prod h_{v_i} p_{v_i}}$	**61.85**	**92.67**	**74.19**	**67.76**
$g_7 = \max h_{v_i} p_{v_i}$	61.15	91.13	73.19	66.61

The active neighbors with different retweet time exert different influence on retweet behaviors.

The majority of pairwise influence is low and a minority is scattered in a fat right tail, thus geometric mean performs better.