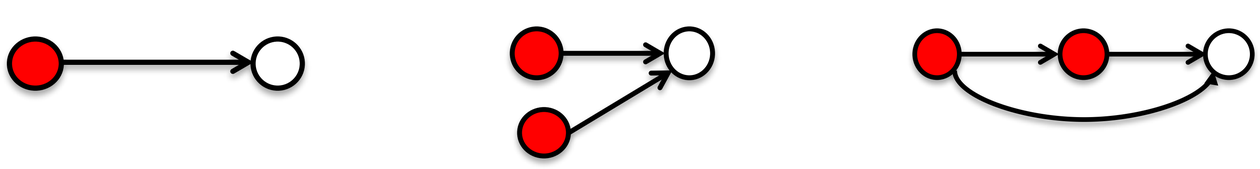
PaperID: 1948

StructInf: Mining Structural Influence from Social Streams

Jing Zhang, Jie Tang⁺, Yuanyi Zhong⁺, Yuchen Mo⁺, Juanzi Li⁺,*
Guojie Song#, Wendy Hall▫, and Jimeng Sun△

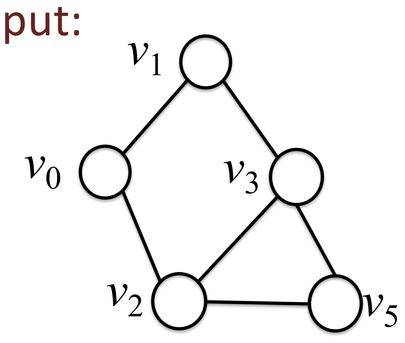
*Renmin University of China, ⁺Tsinghua University,
#Peking University, ▫University of Southampton, △ Georgia Institute of Technology



Question: In which structures, the target nodes are most likely to be activated?

Problem Formulation

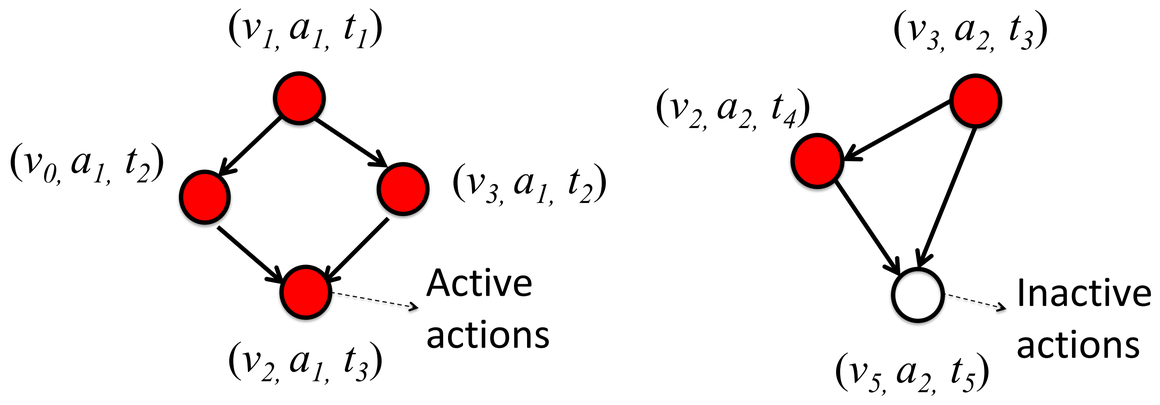
Input:



Network

(v_1, a_1, t_1)
(v_0, a_1, t_2)
(v_3, a_1, t_2)
(v_2, a_1, t_3)
(v_3, a_2, t_3)
(v_2, a_2, t_4)
...

Streaming actions



Action diffusion graph

Output:

Influence Probability		$l_t \in L$	$l_t \notin L$
$\mathrm{IP}(C_k) = \frac{x_k}{x_k + y_k}$	C_k	x_k	y_k
	$\overline{C_k}$	z_k	w_k

Structural influence:
Influence Probabilities of a structure C_k that can be found in the action diffusion graph.

Structural Influence Measurement

Basic method

- Maintain a queue and a map to record the diffusion edges within recent time interval.
- To calculate x_k, active actions are newly arrived actions.
- To calculate y_k, inactive actions are actions that are outdated.
- Enumerate structures by extending neighboring actions of active or inactive actions.
- To avoid duplicate enumeration, assign each action an incremental (unique) label when it arrives, and make the labels of the selected actions smaller than those in the candidate actions.

Sampling based methods

- Sampling1 – Randomly sample nodes when enumerating influence patterns.
- Sampling2 – Randomly reserve edges when building diffusion graph.
- Sampling3 – Combine Sampling1 and Sampling2.

They are unbiased sampling methods

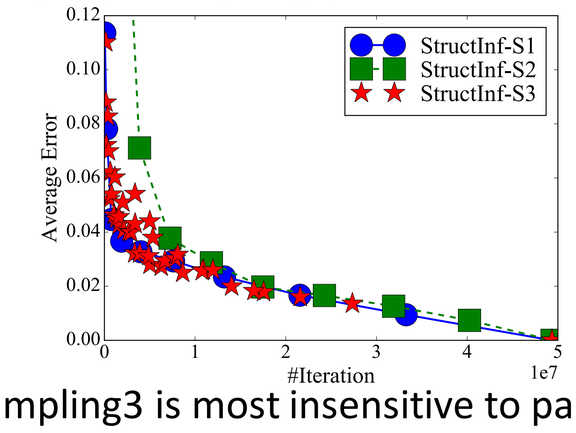
Results

k	C_k	IP_k	$\widetilde{\mathrm{IP}}_k$	U_{IP_k}	k	C_k	IP_k	$\widetilde{\mathrm{IP}}_k$	U_{IP_k}
1		0.066	0.066	0.020	11		0.038	0.038	0.720
2		0.074	0.074	0.085	12		0.186	0.186	0.088
3		0.111	0.110	0.425	13		0.399	0.392	1.785
4		0.307	0.304	0.928	14		0.063	0.062	0.616
5		0.069	0.069	0.530	15		0.619	**0.615**	0.548
6		0.091	0.090	0.358	16		0.444	**0.439**	1.378
7		0.067	0.067	0.236	17		0.070	0.070	0.074
8		0.106	0.099	5.852	18		0.420	**0.416**	0.890
9		0.381	0.388	1.666	19		0.662	**0.645**	2.696
10		0.165	0.162	1.128	20		0.485	**0.479**	1.239

$\widetilde{\mathrm{IP}}_k$: Approximate influence probability

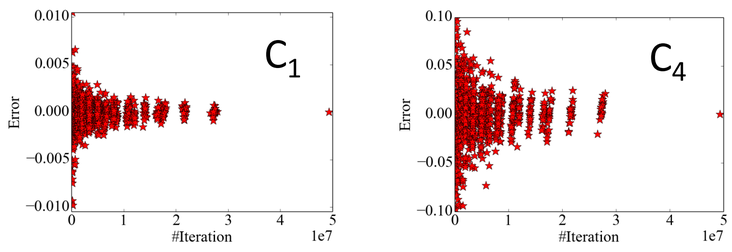
U_{IP_k} : Relative error of approximate values

Trade-off between error and time by varying sampling probabilities:

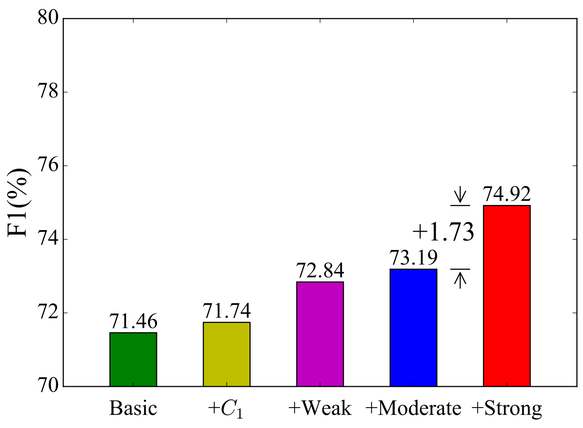


Sampling3 is most insensitive to parameters

Convergence of relative error:



Retweet prediction:



Basic: #friends, gender, status, etc.

C_1 : the number of active neighbors

Weak: $\widetilde{\mathrm{IP}}_k < 0.1$

Moderate: $0.1 \leq \widetilde{\mathrm{IP}}_k < 0.3$

Strong: $\widetilde{\mathrm{IP}}_k > 0.3$