Local Motif Clustering on Time-Evolving Graphs

Presenter: Dongqi Fu
Email: dongqif2@illinois.edu

Dongqi Fu (UIUC)

Dawei Zhou (UIUC)

Jingrui He (UIUC)

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Roadmap

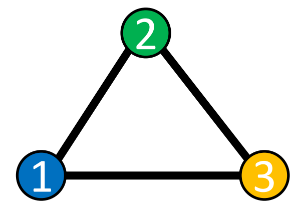
- Motivation
- Problem Definition
- Proposed L-MEGA Framework
- Experiments
- Conclusion

Graph Motifs

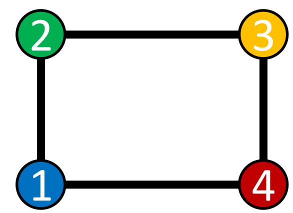
1st-order motif (e.g., node)

2nd-order motif (e.g., edge)

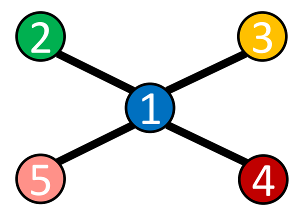
3rd-order motif (e.g., triangle)



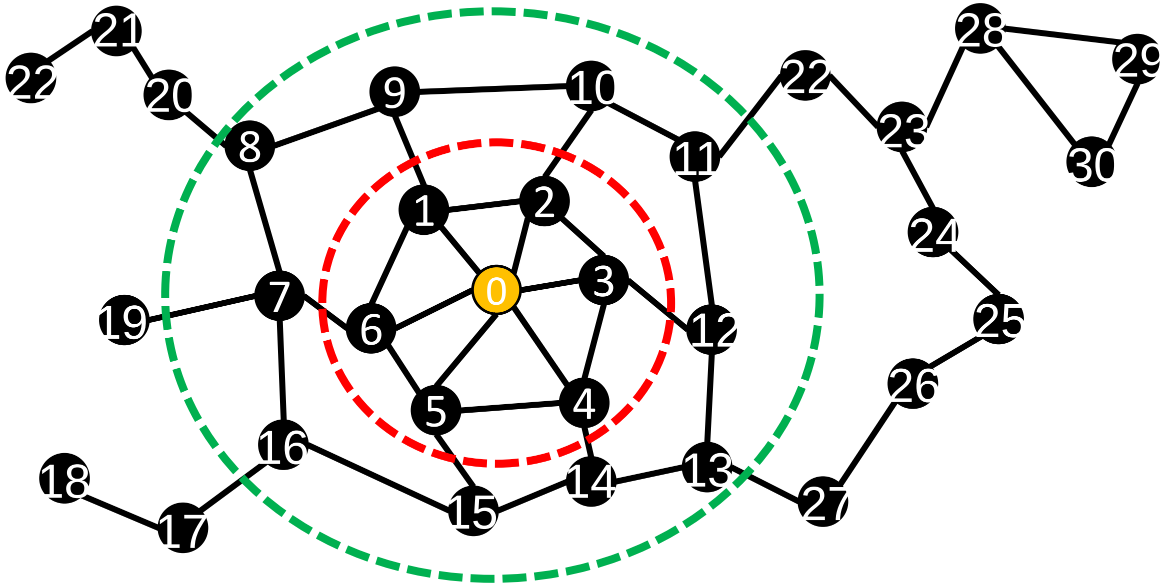
4th-order motif (e.g., loop)



5th-order motif (e.g., star)



Motif Clustering



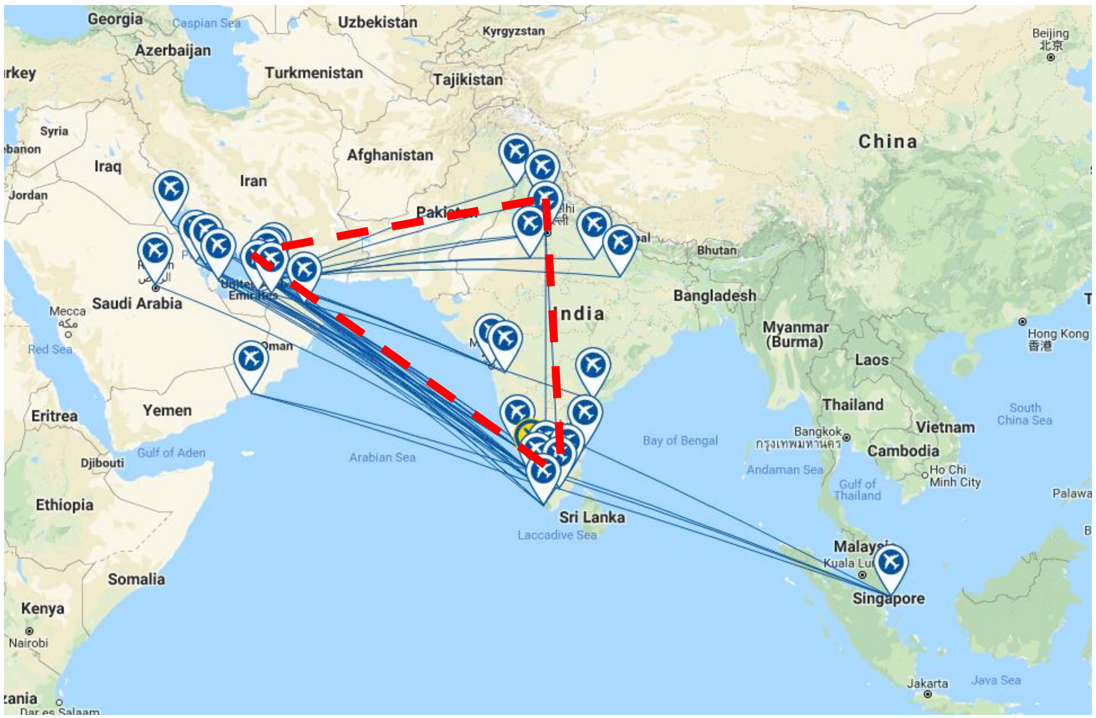
0 Selected Seed Node

Edge-based (2nd-order motif) Clustering

Triangle-based (3rd-order motif) Clustering

Real-World Applications

- In the air traffic network, **triangle motifs** may identify **frequent flight patterns**.



Air Traffic Network

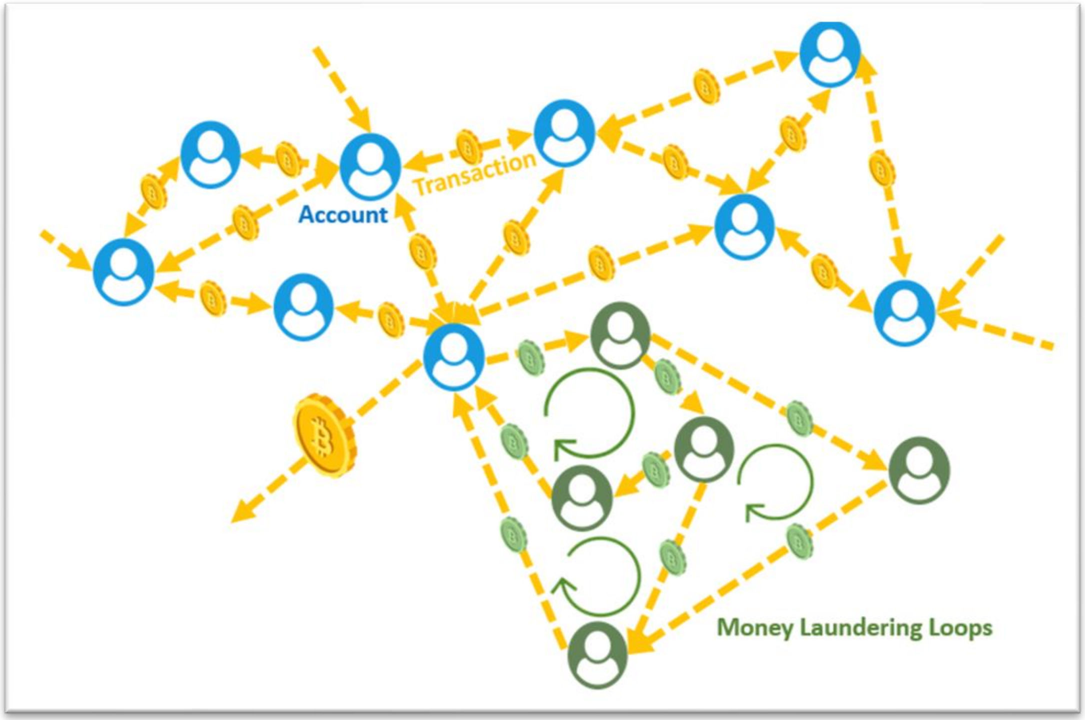
Triangle Motif

Data snapshot: https://blueswandaily.com/

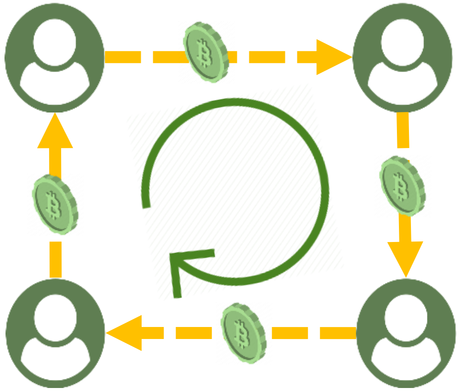
Tao Wu, Austin R. Benson, David F. Gleich: General Tensor Spectral Co-clustering for Higher-Order Data. NeurIPS 2016: 2559-2567

Real-World Applications

- In the online transaction network, **loop motifs** may be associated with the **money laundering activities**.



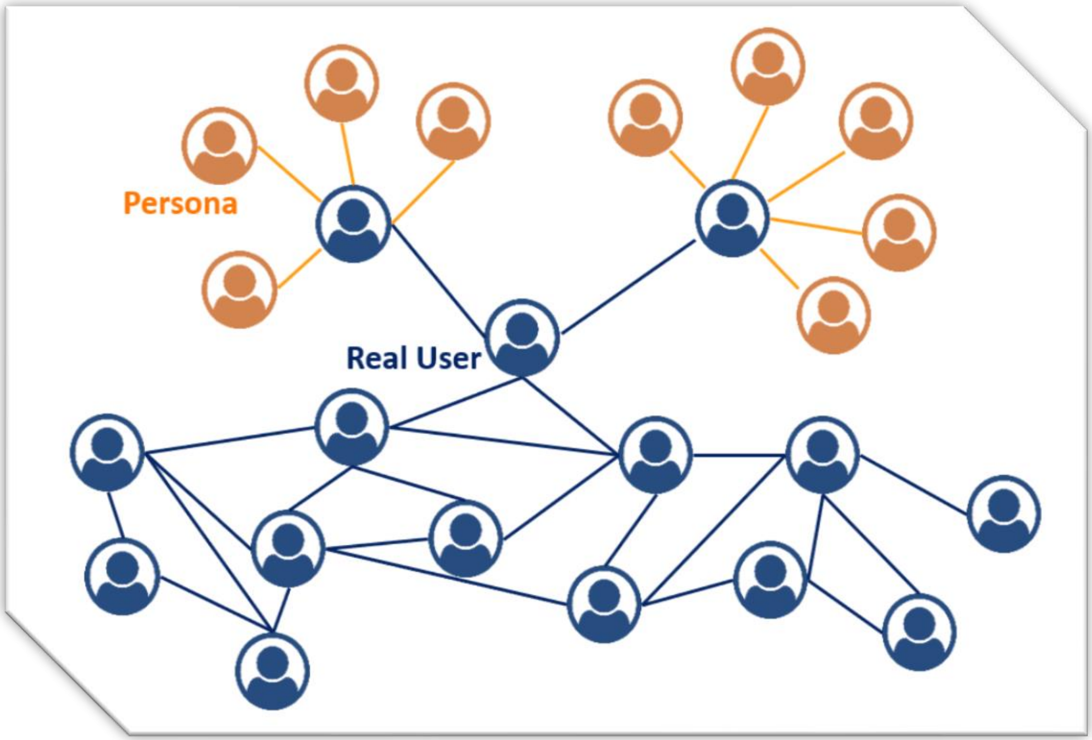
Online Transaction Network



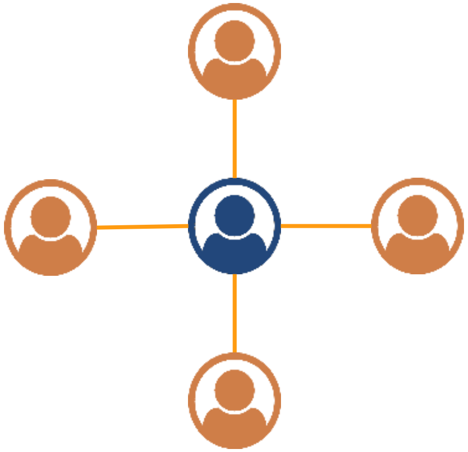
Loop Motif

Real-World Applications

- In the social network, **star motifs** may indicate **personas**.



Social Network

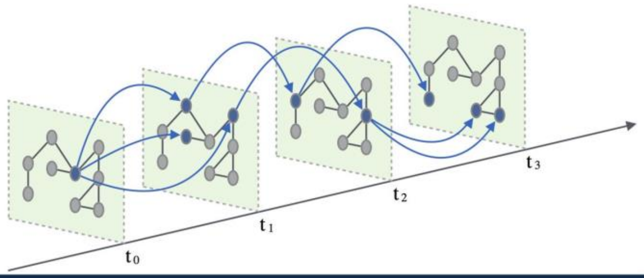


Star Motif

Existing Work

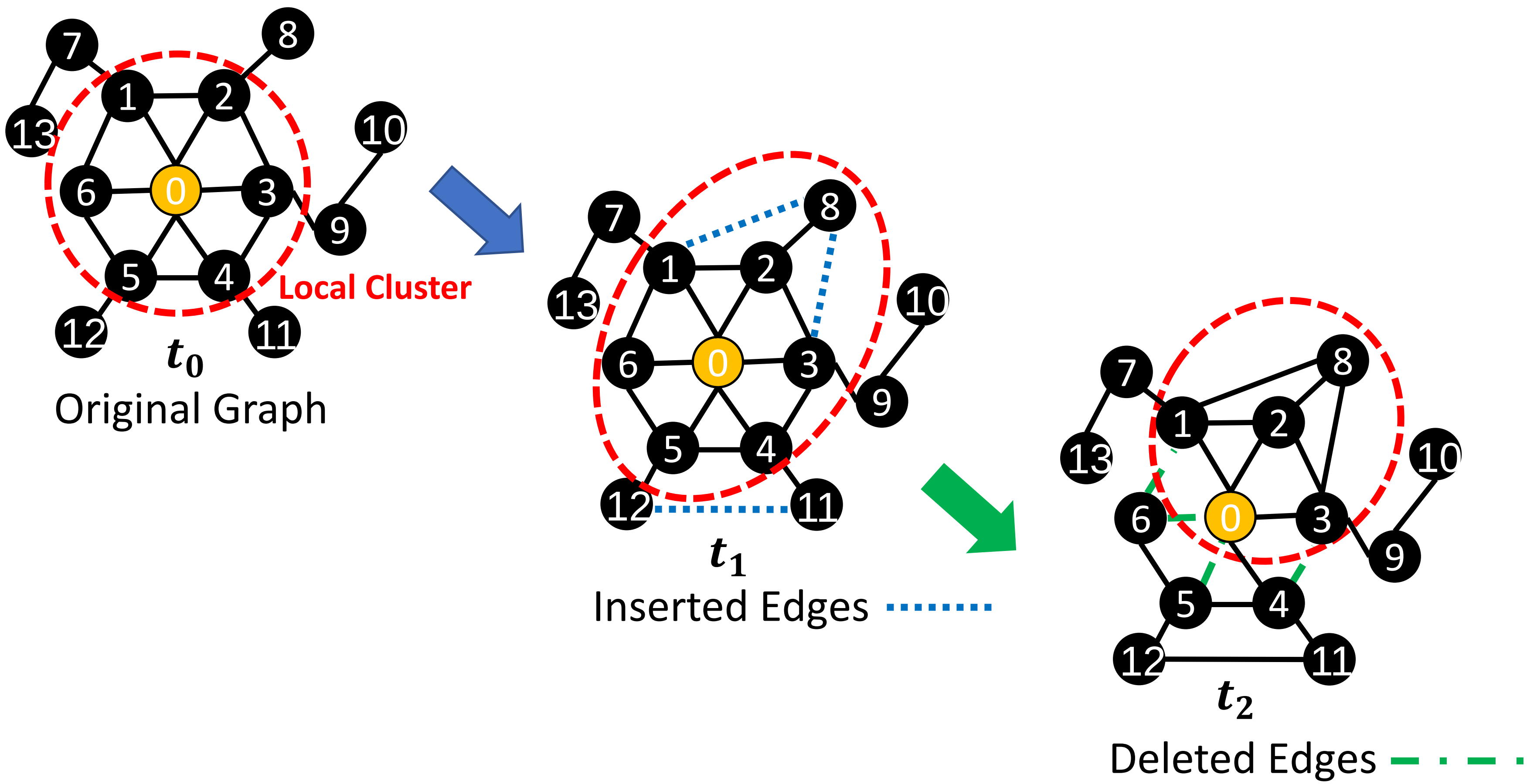
- Existing motif clustering algorithms
  - Local clustering algorithms: HOSPLOC [Zhou et al., 2017], MAPPR [Yin et al., 2017].
  - Spectral clustering algorithms: TSC [Benson et al., 2015], GTSC [Wu et al., 2016], MSC [Benson et al., 2016].
  - Embedding-based algorithms: MCN [Lee et al., 2019], HONE [Rossi et al., 2020].

- Major drawbacks of existing algorithms
  - Designed for **static graphs** and may not capture dynamic patterns.
  - **Computationally prohibitive** on large time-evolving graphs.



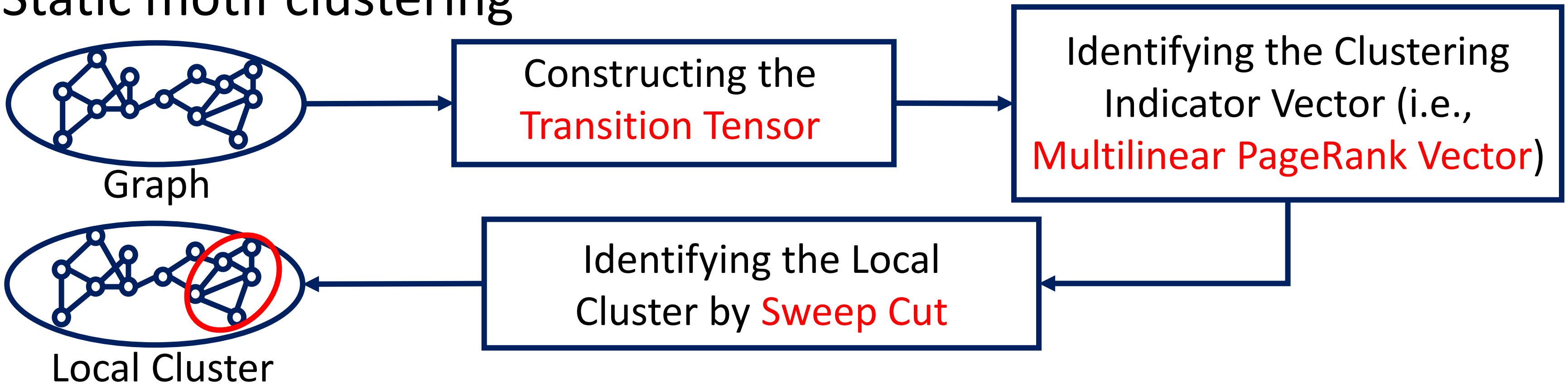
An Illustrative Example

- Local motif clustering on a time-evolving graph



Preliminaries and Challenges

- Static motif clustering



- Challenges (i.e., when the new graph arrives)
  - Challenge 1 (Transition tensor changes): how could we filter some unimportant updated edges before updating the transition tensor?
  - Challenge 2 (Multilinear PageRank vector changes): how could we track multilinear PageRank from last time instead of resolving it?
  - Challenge 3 (Local cluster changes): how could we identify the new local cluster efficiently by leveraging the previous local cluster?

David F. Gleich, Lek-Heng Lim, Yongyang Yu: Multilinear PageRank. SIAM J. Matrix Anal. Appl. 36(4): 1507-1541 (2015)

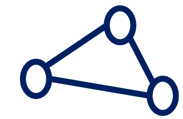
Roadmap

- Motivation
- Problem Definition
- Proposed L-MEGA Framework
- Experiments
- Conclusion

Evaluation Metric

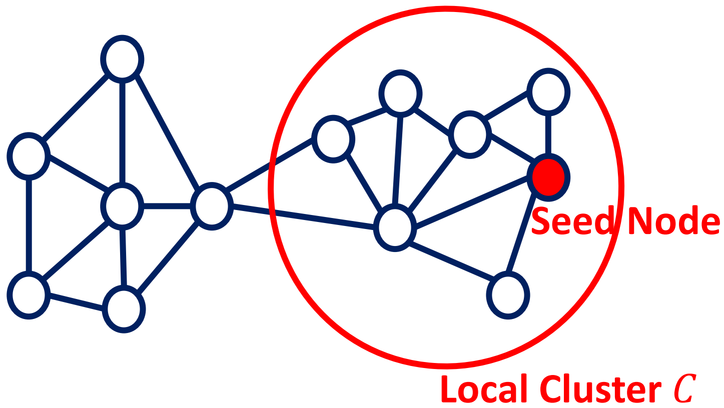
- Motif conductance

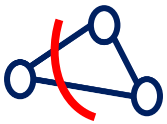
$$\Phi(C, N) = \frac{cut(C, N)}{\min\{\mu(C, N), \mu(\bar{C}, N)\}}$$

→ N: motif structure 

→ $cut(C, N)$: # of motifs being cut

→ $\mu(C, N)$: # of motif end points in C



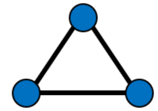
→ $cut(C, N)$: 1 

→ $\mu(C, N)$: 17

→ $\Phi(C, N)$: 1 / 17

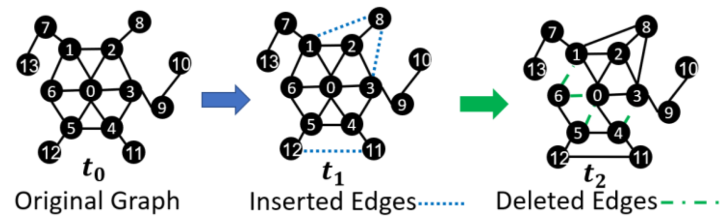
Problem Definition

- Local motif clustering on time-evolving graphs

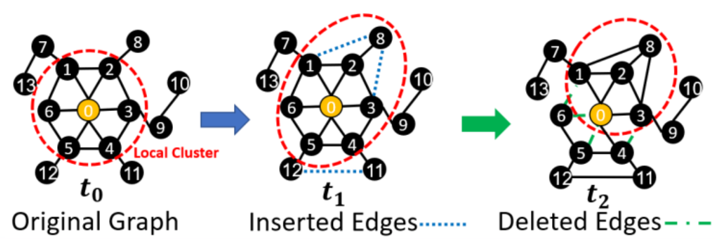
  Given:
  - A user-defined motif structure N; 
  - A sequence of snapshots of the time-evolving graph

$$\tilde{G} = \{G^{(0)}, G^{(0)}, \dots, G^{(T)}\};$$

  - A seed node v; 0
  - A motif conductance upperbound φ.



  Find:
  - A local cluster $C^{(t)}$ near seed node v such that $\Phi\left(C^{(t)}, N\right) \le \varphi$ at each time stamp $t \in \{1, 2, \dots, T\}$.

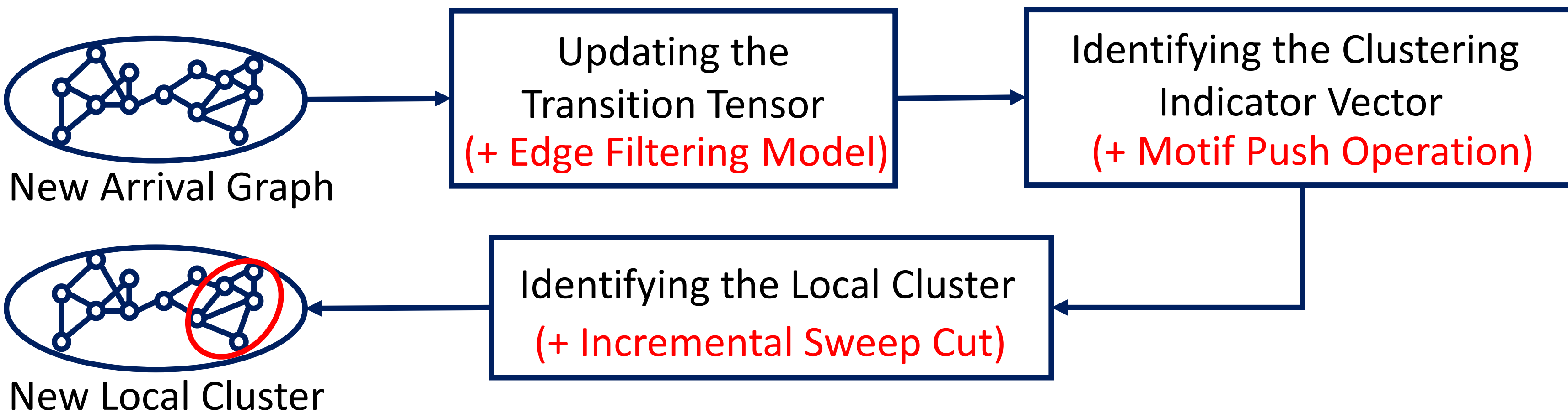


Roadmap

- **Motivation**
- **Problem Definition**
- **Proposed L-MEGA Framework**
- **Experiments**
- **Conclusion**

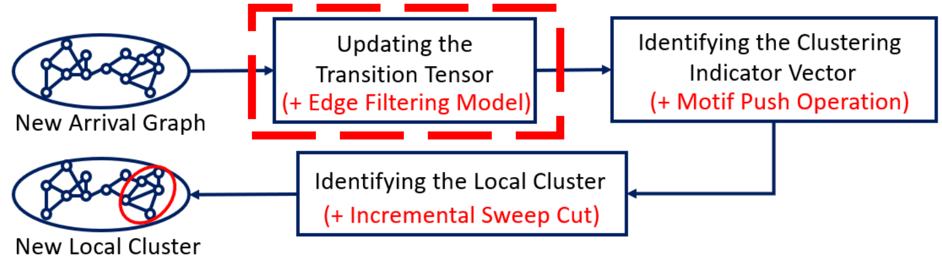
Overview of L-MEGA

- L-MEGA (**L**ocal **M**otif Clustering on Time-**E**volving **G**r**a**phs)
  - Identifies the evolution pattern of the local motif cluster ***effectively*** and ***efficiently***.
- Key idea of L-MEGA
  - Tracks the local motif cluster via three speed-up techniques for three mentioned challenges.



L-MEGA Framework

L-MEGA Framework

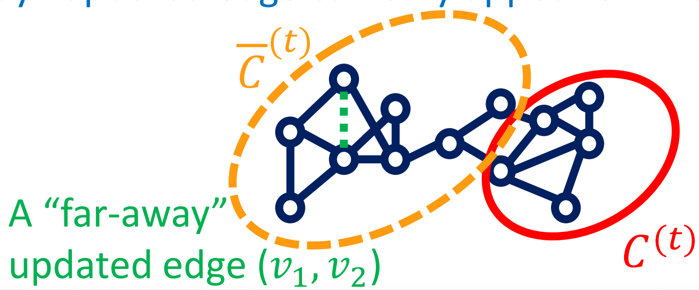


- Edge filtering model
  - Intuition: Some "far-away" updated edges at time t will not influence the evolution of the previously identified local cluster at time t+1.
  - Solution: Identify "far-away" edge (v_1, v_2), filter it out before updating the transition tensor, and save it for the future.

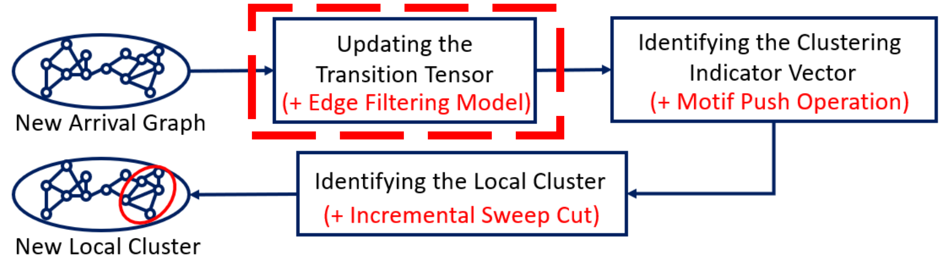
$\pi(j+1)$: the node on $\overline{C}^{(t)}$ with the maximum probability mass

Condition 1: $$\frac{\boldsymbol{x}^{(t)}(v_1)}{\boldsymbol{d}^{(t)}(v_1)} < \frac{\boldsymbol{x}^{(t)}(\pi(j+1))}{\boldsymbol{d}^{(t)}(\pi(j+1))}, \frac{\boldsymbol{x}^{(t)}(v_2)}{\boldsymbol{d}^{(t)}(v_2)} < \frac{\boldsymbol{x}^{(t)}(\pi(j+1))}{\boldsymbol{d}^{(t)}(\pi(j+1))}$$

"far-away" updated edge can only appear on the complement $\overline{C}^{(t)}$



L-MEGA Framework



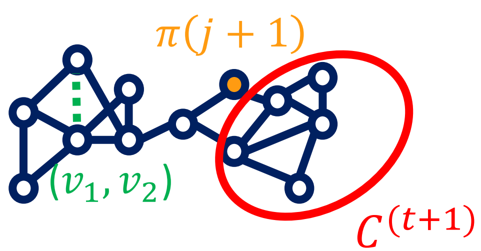
- Edge filtering model
  - Intuition: Some "far-away" updated edges at time t will not influence the evolution of the previously identified local cluster at time t+1.
  - Solution: Identify "far-away" edge (v_1, v_2), filter it out before updating the transition tensor, and save it for the future.

γ: the maximum probability mass contribution from updated edge (v_1, v_2)

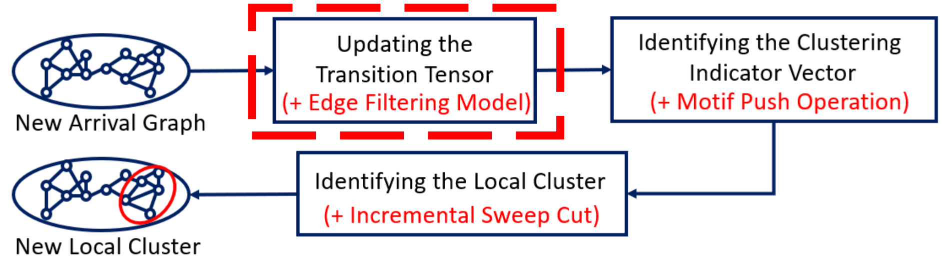
$\pi(j+1)$: the node on $\overline{C}^{(t)}$ with the maximum probability mass

Condition 2: $$\frac{\gamma + \boldsymbol{x}^{(t)}(\pi(j+1))}{\boldsymbol{d}^{(t)}(\pi(j+1))} < \frac{\boldsymbol{x}^{(t)}(\pi(j))}{\boldsymbol{d}^{(t)}(\pi(j))}$$

"far-away" updated edge cannot send node $\pi(j+1)$ to the local cluster at time t+1



L-MEGA Framework

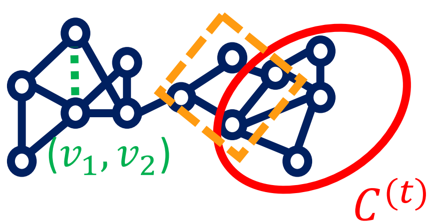


- Edge filtering model
  - Intuition: Some “far-away” updated edges at time t will not influence the evolution of the previously identified local cluster at time t+1.
  - Solution: Identify “far-away” edge (v_1, v_2), filter it out before updating the transition tensor, and save it for the future.

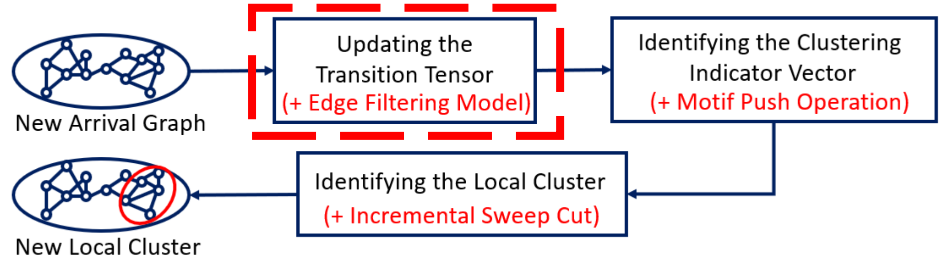
D_{ist} denotes the shortest distance from v_1 to any node of $C^{(t)}$

Condition 3: $D_{ist}\left(v_1, C^{(t)}\right) > k-1,\ D_{ist}\left(v_2, C^{(t)}\right) > k-1$

“far-away” updated edge cannot induce any new motifs involved in the previous graph cut



L-MEGA Framework



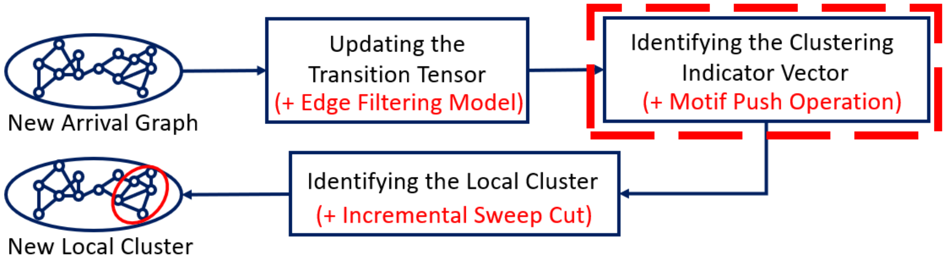
- Edge filtering model
  - Intuition: Some “far-away” updated edges at time t will not influence the evolution of the previously identified local cluster at time t+1.
  - Solution: Identify “far-away” edge (v_1, v_2), filter it out before updating the transition tensor, and save it for the future.
  - Time Complexity:

$$O(|V^{(t+1)}|^{k-2}) \xrightarrow[\text{to}]{\text{reduced}} O(|V^{(t+1)}| + |E^{(t+1)}| \log |V^{(t+1)}|)$$

updating transition tensor for one updated edge

filtering one “far-away” updated edge

L-MEGA Framework



- Motif push operation
  - Intuition: Resolving multilinear PageRank is time-consuming, we may track it from last time instead of resolving it.
  - Solution: Track multilinear PageRank $\boldsymbol{x}^{(t+1)}$ from $\boldsymbol{x}^{(t)}$.

$$\boldsymbol{Appr}(\boldsymbol{x}^{(t+1)}) = \boldsymbol{x}^{(t)}$$

moving the largest entry $\boldsymbol{r}^{(t+1)}(i)$ coordinately

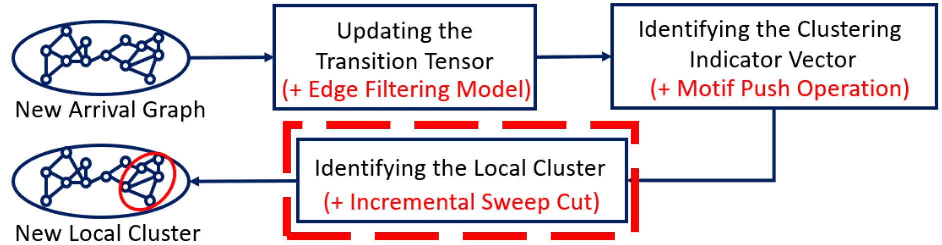
$\boldsymbol{r}^{(t+1)}$: divergence between $\boldsymbol{Appr}(\boldsymbol{x}^{(t+1)})$ and $\boldsymbol{x}^{(t+1)}$

$$\boldsymbol{r}^{(t+1)} + \boldsymbol{Appr}(\boldsymbol{x}^{(t+1)}) = \boldsymbol{x}^{(t+1)}$$

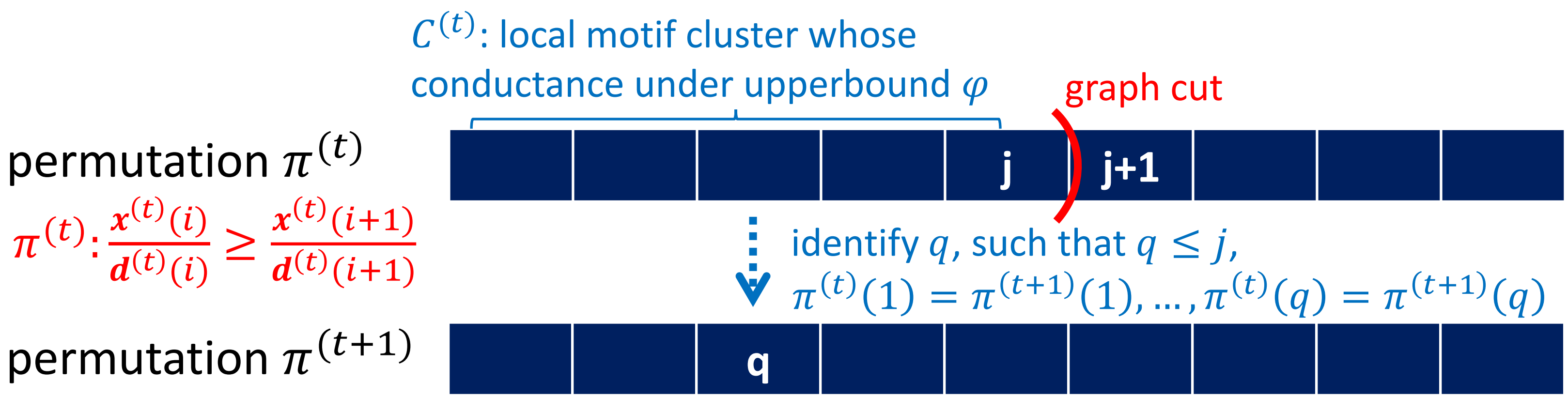
adding back the additional divergence due to the moving

  - Time Complexity: $O(|V^{(t+1)}|^k) \xrightarrow[\text{to}]{\text{reduced}} O(\frac{1}{\xi^k})$, where $\xi \propto \frac{1}{\log_2 \mu(V^{(t+1)})}$

L-MEGA Framework



- Incremental sweep cut
  - Intuition: If updated edge set only contains inserted edges on complement $\overline{C}^{(t)}$ after edge filtering and $\mu\left(C^{(t)}, N\right) < \mu(\bar{C}^{(t)}, N)$, then $|C^{(t+1)}| \geq |C^{(t)}|$.
  - Solution: Identify the shared sequence between two consecutive time permutations and start from the first different entry.



  - Time Complexity: Reduces q iterations, each iteration costing $O(|V^{(t+1)}|^2)$.

Roadmap

- Motivation
- Problem Definition
- Proposed L-MEGA Framework
- Experiments
- Conclusion

Experimental Setup

- Comparison methods
  - Static edge-based clustering algorithms: Nibble [Spielman and Teng, 2013].
  - Dynamic edge-based clustering algorithms: ISC [Ning et al., 2010], TPPR[Ohsaka et al., 2015].
  - Static motif-based clustering algorithms: HOSPLOC [Zhou et al., 2017], MAPPR [Yin et al., 2017].
  - Dynamic motif-based clustering algorithms: Our L-MEGA, L-MEGA-1 (without the edge filtering model), L-MEGA-2 (without the motif push operation), L-MEGA-3 (without the incremental sweep cut).

- Dataset

Network	Category	$\|V\|$	$\|E\|$	Time Span
Alpha	Rating	3,783	14,124	62 months
OTC	Rating	5,881	21,492	62 months
Call	Communication	6,809	7,967	4 months
Contact	Interaction	10,972	44,517	3 months

Effectiveness Comparison

- L-MEGA outperforms baseline methods in terms of four metrics
  - Conductance (the lower the better).
  - Third-order conductance (the lower the better).
  - Triangle density (the higher the better).
  - Time consumption (the lower the better).

Methods	Alpha			
	conductance	third-order conductance	triangle density	time
Nibble	0.4909 ± 0.0060	0.4555 ± 0.0454	0.2355 ± 0.1033	18.4073 ± 5.9853
TPPR	0.4923 ± 0.0089	0.4994 ± 0.1188	0.1613 ± 0.0934	12.4094 ± 5.7653
ISC	$\boxed{0.3334 \pm 0.0000}$	1.0000 ± 0.0000	0.0000 ± 0.0000	56.6376 ± 0.0000
MAPPR	0.4947 ± 0.0008	0.5852 ± 0.0104	0.0712 ± 0.0030	43.0597 ± 2.9107
HOSPLOC	0.4915 ± 0.0080	0.4816 ± 0.0576	0.1891 ± 0.0859	237.6121 ± 12.5513
L-MEGA	0.4712 ± 0.0586	$\boxed{\mathbf{0.4097 \pm 0.0278}}$	0.2561 ± 0.1008	$\boxed{\mathbf{8.2032 \pm 4.8534}}$
L-MEGA-1	0.4728 ± 0.0102	0.4676 ± 0.0344	0.2490 ± 0.0736	241.4762 ± 13.3320
L-MEGA-2	0.4944 ± 0.0036	0.4369 ± 0.0428	$\boxed{0.3819 \pm 0.0737}$	14.8578 ± 4.0788
L-MEGA-3	0.4712 ± 0.0586	0.4097 ± 0.0278	0.2561 ± 0.1008	11.4955 ± 4.2939

Performance on Alpha Network at the Last Time Stamp

Effectiveness Comparison

Methods	Alpha				OTC			
	conductance	third-order conductance	triangle density	time	conductance	third-order conductance	triangle density	time
Nibble	0.4909 ± 0.0060	0.4555 ± 0.0454	0.2355 ± 0.1033	18.4073 ± 5.9853	0.4963 ± 0.0045	0.5091 ± 0.0941	0.1582 ± 0.1076	63.1869 ± 34.2154
TPPR	0.4923 ± 0.0089	0.4994 ± 0.1188	0.1613 ± 0.0934	12.4094 ± 5.765[illegible]	[illegible]21	0.5751 ± 0.1106	0.1524 ± 0.1320	39.1307 ± 19.3550
ISC	0.3334 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	56.6376 ± 0.0[illegible]	[illegible]	0.5656 ± 0.0000	0.1908 ± 0.0000	195.5490 ± 0.0000
MAPPR	0.4947 ± 0.0008	0.5852 ± 0.0104	0.0712 ± 0.0030	43.0597 ± 2.91[illegible]	[illegible]	0.5404 ± 0.0023	0.0904 ± 0.0001	207.5004 ± 1.1757
HOSPLOC	0.4915 ± 0.0080	0.4816 ± 0.0576	0.1891 ± 0.0859	237.6121 ± 12.5513	[illegible] ± 0.0041	0.5080 ± 0.0722	0.2000 ± 0.1172	753.3742 ± 51.6812
L-MEGA	0.4712 ± 0.0586	**0.4097 ± 0.0278**	0.2561 ± 0.1008	**8.2032 ± 4.8534**	0.4652 ± 0.0074	**0.4102 ± 0.0620**	0.2946 ± 0.0719	**32.4308 ± 46.8278**
L-MEGA-1	0.4728 ± 0.0102	0.4676 ± 0.0344	0.2490 ± 0.0736	241.4762 ± 13.3320	0.4733 ± 0.0074	0.4622 ± 0.0547	0.2578 ± 0.0961	778.5583 ± 33.4156
L-MEGA-2	0.4944 ± 0.0036	0.4369 ± 0.0428	0.3819 ± 0.0737	14.8578 ± 4.0788	0.4860 ± 0.0013	0.4750 ± 0.0175	0.5318 ± 0.0141	32.6827 ± 1.6759
L-MEGA-3	0.4712 ± 0.0586	0.4097 ± 0.0278	0.2561 ± 0.1008	11.4955 ± 4.2939	0.4652 ± 0.0074	0.4102 ± 0.0620	0.2946 ± 0.0719	45.5937 ± 45.6706

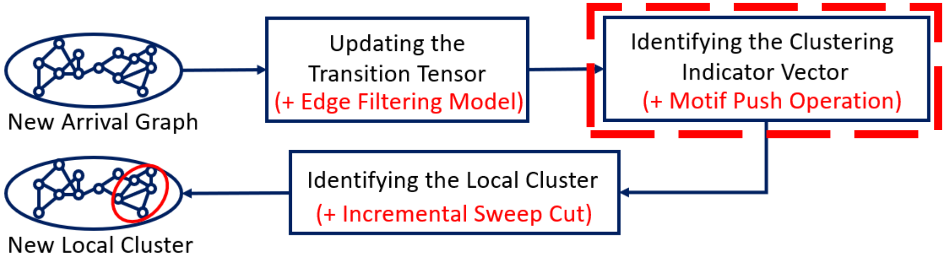
8.2032s

Methods	Call				Contact			
	conductance	third-order conductance	triangle density	time	conductance	third-order conductance	triangle density	time
Nibble	0.0792 ± 0.0309	0.5675 ± 0.4809	0.0249 ± 0.0384	13.5155 ± 2.7236	0.3536 ± 0.0925	0.2878 ± 0.1857	0.0017 ± 0.0015	33.7139 ± 0.1147
TPPR	0.1910 ± 0.1399	0.5589 ± 0.4442	0.0274 ± 0.0420	8.3268 ± 2.160[illegible]	[illegible]23	0.2221 ± 0.1382	0.0025 ± 0.0023	25.0759 ± 0.1416
ISC	0.5893 ± 0.0000	0.5270 ± 0.0000	0.1700 ± 0.0000	27.3982 ± 0.[illegible]	[illegible]	0.5252 ± 0.0000	0.0035 ± 0.0000	1351.1732 ± 0.0000
MAPPR	0.5957 ± 0.0042	0.4401 ± 0.0291	0.2219 ± 0.1869	2938.3853 ± 81.[illegible]	[illegible]3	0.2790 ± 0.0753	0.0006 ± 0.0003	88.6153 ± 0.2981
HOSPLOC	0.1652 ± 0.0485	0.2981 ± 0.3721	0.0296 ± 0.0416	768.4879 ± 1.1554	[illegible] ± 0.1346	0.2308 ± 0.1559	0.0034 ± 0.0051	3443.8829 ± 0.2193
L-MEGA	0.1542 ± 0.0544	0.2866 ± 0.3823	0.0395 ± 0.0448	**1.1316 ± 0.9816**	0.2438 ± 0.1676	**0.1614 ± 0.1443**	0.0042 ± 0.0052	**3.4496 ± 3.0660**
L-MEGA-1	0.1542 ± 0.0544	0.2866 ± 0.3823	0.0395 ± 0.0448	754.2537 ± 0.8341	0.2028 ± 0.1349	0.2172 ± 0.1921	0.0004 ± 0.0005	3492.1936 ± 3.2966
L-MEGA-2	0.2333 ± 0.1839	**0.2713 ± 0.3883**	0.0239 ± 0.0180	17.2451 ± 0.9250	0.2511 ± 0.1121	0.2166 ± 0.1304	0.0003 ± 0.0004	48.1245 ± 1.1528
L-MEGA-3	0.1542 ± 0.0544	0.2866 ± 0.3823	0.0395 ± 0.0448	1.2502 ± 1.0725	0.2438 ± 0.1676	0.1614 ± 0.1443	0.0042 ± 0.0052	3.5800 ± 3.1314

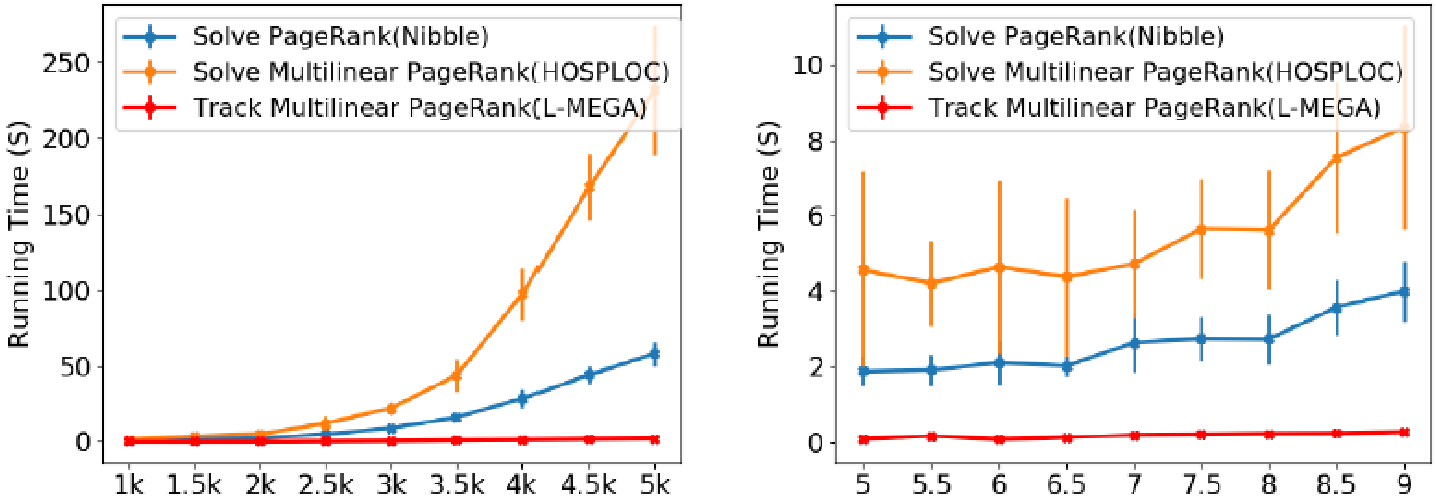
1.1316s

Performance on Four Networks at the Last Time Stamp

Scalability Analysis



- Scalability of motif push operation
  - L-MEGA (motif push operation) is near constant with very slow increase.



(a) The number of vertices

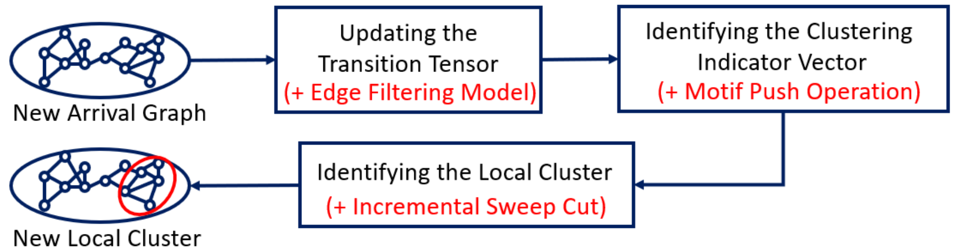
(b) The edge density

Roadmap

- Motivation
- Problem Definition
- Proposed L-MEGA Framework
- Experiments
- Conclusion

Conclusion

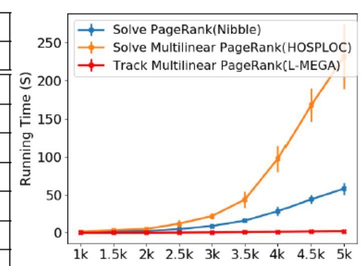
- L-MEGA: A novel local motif clustering framework
  - Edge filtering model.
  - Motif push operation.
  - Incremental sweep cut.



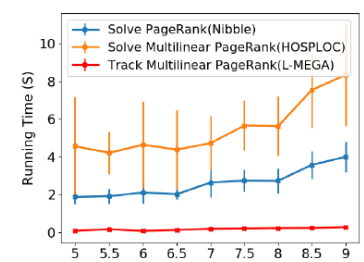
L-MEGA Framework

- Results
  - L-MEGA outperforms baseline methods in finding dense motif clusters and consuming less time on time-evolving graphs.

Methods	Alpha			
	conductance	third-order conductance	triangle density	time
Nibble	0.4909 ± 0.0060	0.4555 ± 0.0454	0.2355 ± 0.1033	18.4073 ± 5.9853
TPPR	0.4923 ± 0.0089	0.4994 ± 0.1188	0.1613 ± 0.0934	12.4094 ± 5.7653
ISC	0.3334 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	56.6376 ± 0.0000
MAPPR	0.4947 ± 0.0008	0.5852 ± 0.0104	0.0712 ± 0.0030	43.0597 ± 2.9107
HOSPLOC	0.4915 ± 0.0080	0.4816 ± 0.0576	0.1891 ± 0.0859	237.6121 ± 12.5513
L-MEGA	0.4712 ± 0.0586	$\mathbf{0.4097 \pm 0.0278}$	0.2561 ± 0.1008	$\mathbf{8.2032 \pm 4.8534}$
L-MEGA-1	0.4728 ± 0.0102	0.4676 ± 0.0344	0.2490 ± 0.0736	241.4762 ± 13.3320
L-MEGA-2	0.4944 ± 0.0036	0.4369 ± 0.0428	0.3819 ± 0.0737	14.8578 ± 4.0788
L-MEGA-3	0.4712 ± 0.0586	0.4097 ± 0.0278	0.2561 ± 0.1008	11.4955 ± 4.2939



(a) The number of vertices



(b) The edge density

Thanks！

Dongqi Fu
(UIUC)

Dawei Zhou
(UIUC)

Jingrui He
(UIUC)

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Please refer to our paper and code at
https://github.com/DongqiFu/L-MEGA