ComPAS: Community Preserving Sampling for Streaming Graphs

Sandipan Sikdar

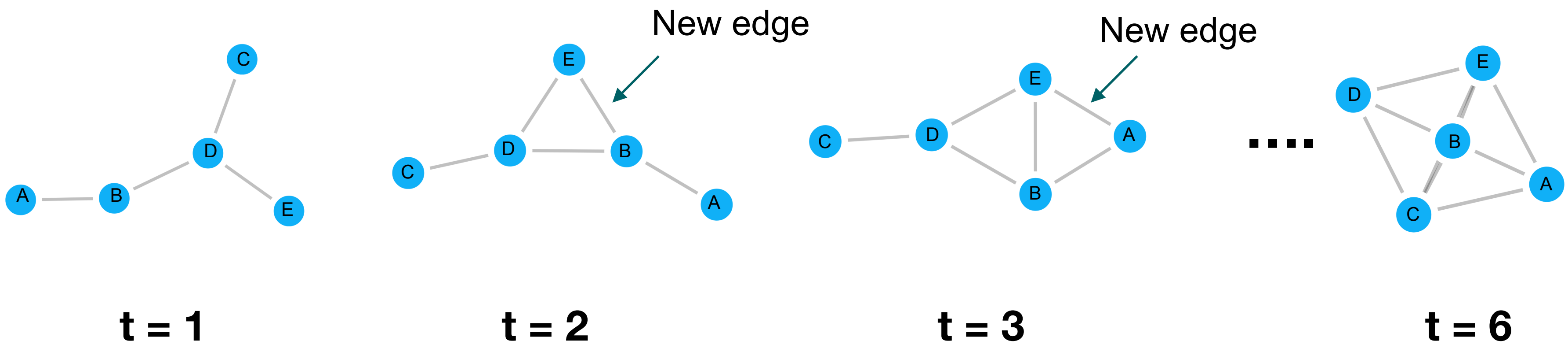
Chair for Computational Social Science and Humanities,

RWTH Aachen

Ref: S. Sikdar, T. Chakraborty, S. Sarkar, N. Ganguly, A. Mukherjee:
ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Streaming Graphs

- Sequence of edges ordered in time
- Graph G is the aggregation of all the edges over time
- Typical examples include citation network, email log, facebook posts



New edge

New edge

t = 1

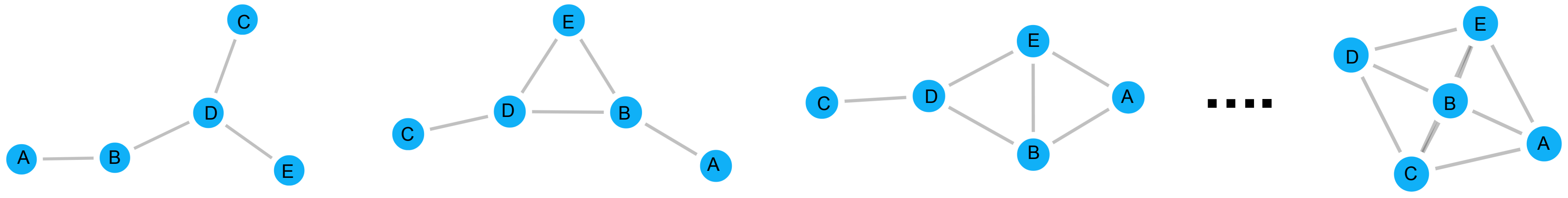
t = 2

t = 3

t = 6

Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Streaming Graph Sampling

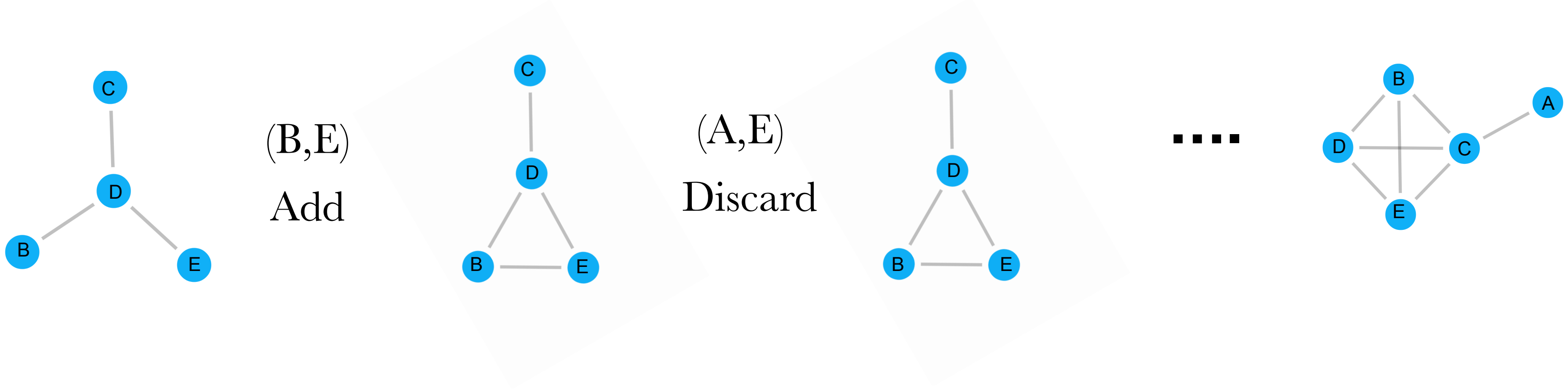


t = 1

t = 2

t = 3

t = 6



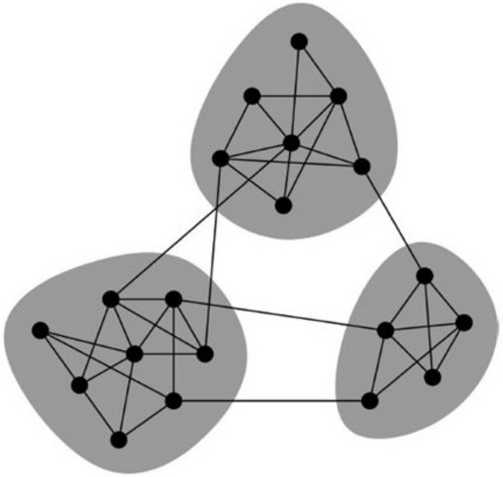
(B,E)
Add

(A,E)
Discard

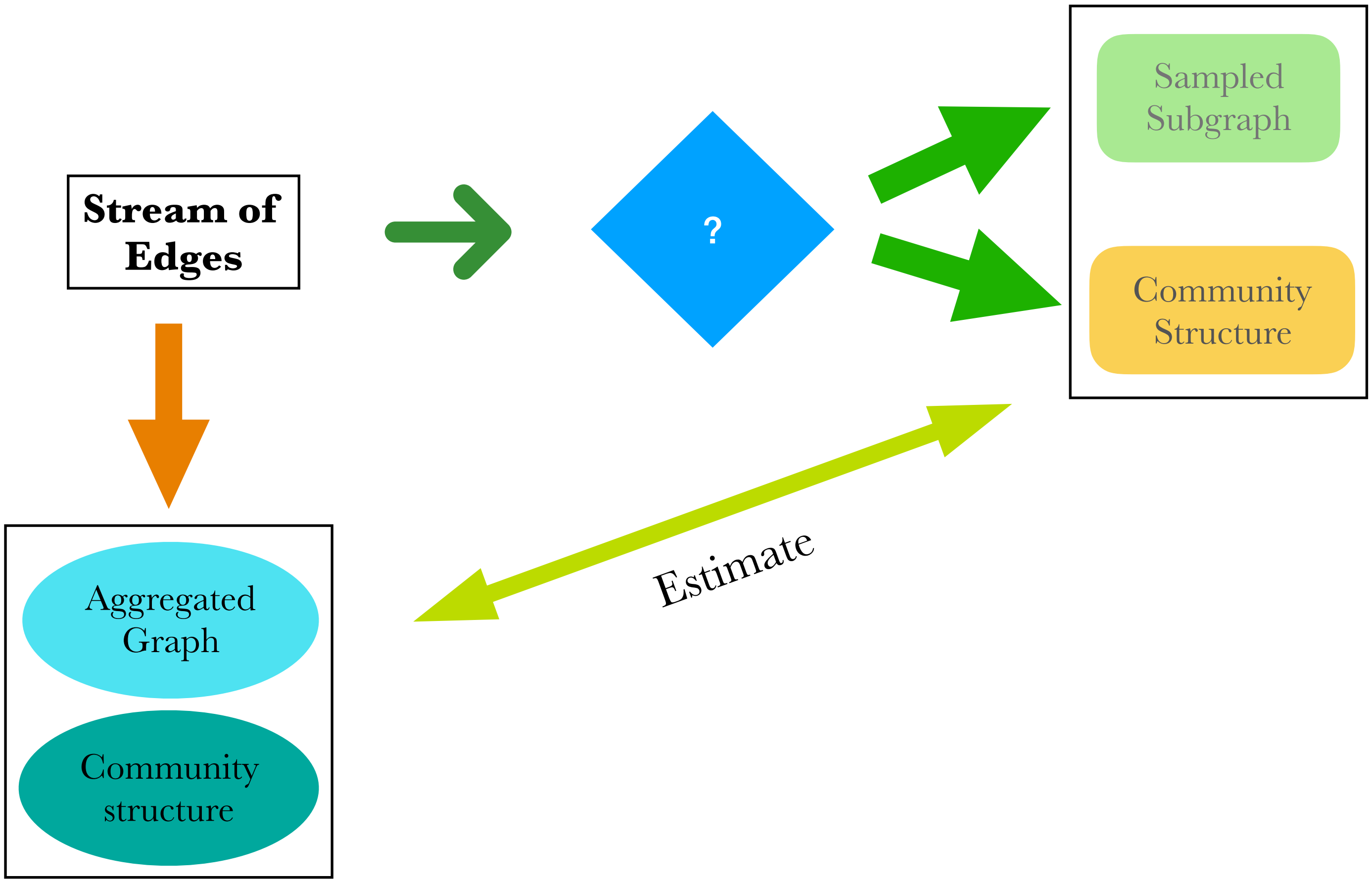
Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Streaming Graph Sampling with Community

- Given a streaming graph G, the objective is obtain a sample G_s such that the properties of G are maintained in G_s
- Existing algorithms are designed for preserving simple structural properties
- We propose ComPAS which is capable of retaining the underlying community structure
- Applications - Obtaining stratified samples in online learning



Sampling Problem



Proposed Algorithm: ComPAS

- Maximize modularity
- Identify high fidelity nodes over time
- Allow merging, splitting and creation of new communities

Proposed Algorithm: ComPAS

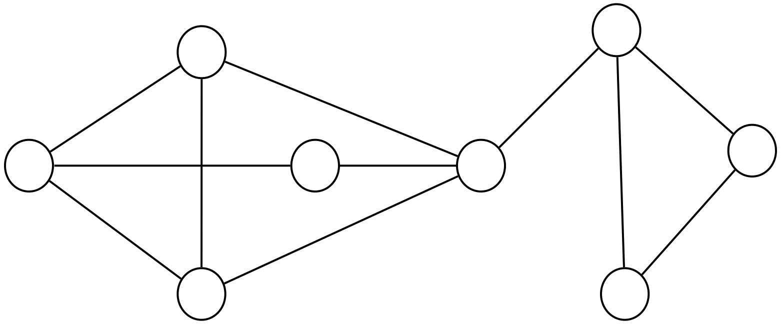
Parameters:

- sample size (n)
- alpha (0< α <1)
- Buffer (H) consisting of two variables
  - H_c - Number of times a node is encountered
  - H_p - Current parent

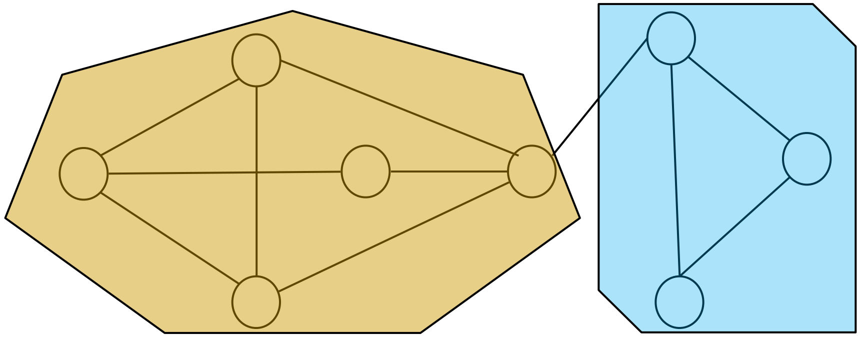
Node	Count	Parent
i	1	d
j	3	l
k	1	m
l	4	j
m	3	e
n	1	k

Dynamics of ComPAS

- Keep adding edges into the sample as long as a certain number of nodes are inserted ($\alpha * n$)



- Once the threshold is reached a pre-selected community detection algorithm is executed on the sample to obtain initial community structure.



Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

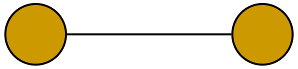
Role of Buffer

- From this point on whenever a new node is encountered it is pushed to buffer
- Estimate the importance of a node
- More recurrent node is perhaps more important

Node	Count	Parent
i	1	d
j	3	l
k	1	m
l	4	j
m	3	e
n	1	k

Position of Nodes

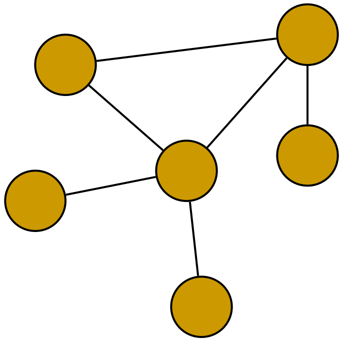
New



In Buffer

Node	Count	Parent
i	1	d
j	3	l
k	1	s
l	4	j
m	3	e
n	1	k

In Sample



Genesis of Six Modules

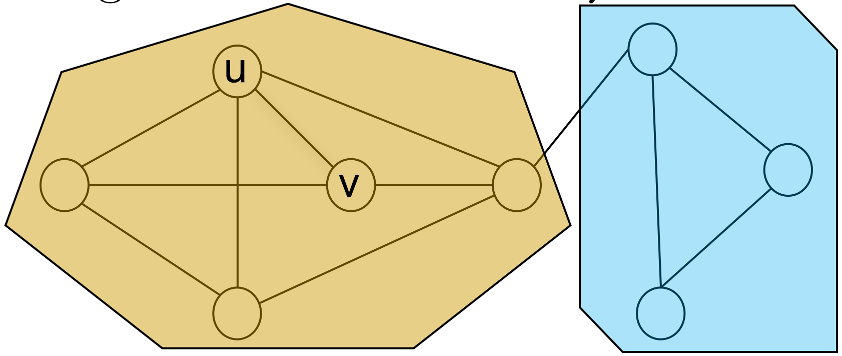
- Both vertices in the sample
- Both vertices in buffer
- One in sample and one in buffer
- One in sample and one is new
- One in buffer and one is new
- Both are new

- Constraints
  - A new node cannot be directly added to the sample
  - Only nodes from buffer are eligible to enter the sample
  - If sample size is reached node must be deleted to make way

Both in Sample

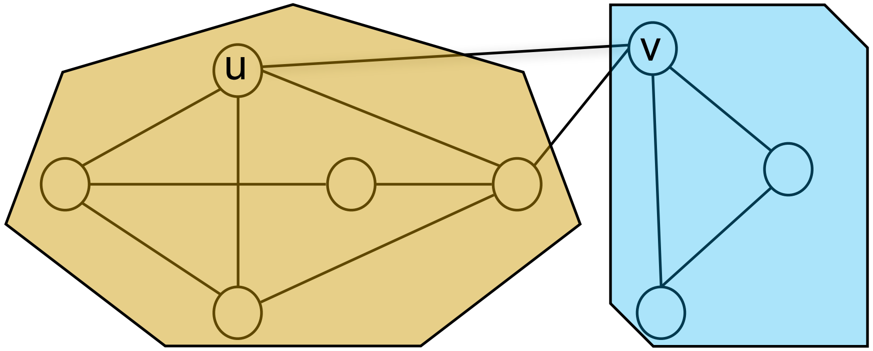
This can be further divided into two sub cases -

- The edge is intra-community



Add the edge to the sample

- The edge is inter-community



- u may leave its current community and join v's
- v may leave its current community and join u's
- u and v leave their current communities and form new one

Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Both in Buffer

- edge (j,k)

Node	Count	Parent
i	1	d
j	3	l
k	1	m
l	4	j
m	3	e
n	1	k

Node	Count	Parent
i	1	d
j	**4**	l
k	**2**	m
l	4	j
m	3	e
n	1	k

One in Sample one in Buffer

- edge (u,k)

Node	Count	Parent
i	1	d
j	4	l
k	2	m
l	4	j
m	3	e
n	1	k

Node	Count	Parent
i	1	d
j	4	l
k	**3**	m
l	4	j
m	3	e
n	1	k

Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Dynamics of ComPAS

✓ Both vertices in the sample

✓ Both vertices in buffer

✓ One in sample and one in buffer

✗ One in sample and one is new

✗ One in buffer and one is new

✗ Both are new

Dynamics of ComPAS

- ✓ Both vertices in the sample
- ✓ Both vertices in buffer
- ✓ One in sample and one in buffer
- ✗ One in sample and one is new
- ✗ One in buffer and one is new
- ✗ Both are new

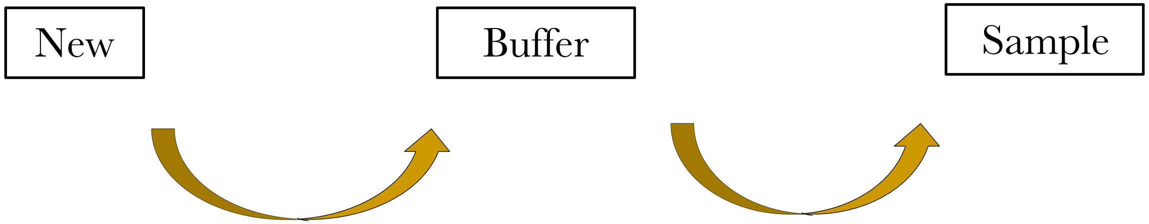
At least one node is new

Entry of a new Node

- In the subsequent cases at least one node is new
- This node triggers rearrangement -
  - Remove node from buffer to make way for new node

    Preferentially (based on $H_c(x)$) remove a node x from buffer with additional constraint that P(x) in sample
  - Remove node from sample to make way for x

    Node with lowest degree and clustering coefficient is removed from sample



Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Deletion of a Node from Sample

- New node (v) is encountered
- Buffer is full
- Sample size has been reached
    1. Preferentially select u from buffer and add it to sample
    2. Assign u the community of its parent P(u)
    3. Remove a node w with the lowest degree and clustering coefficient from sample
    4. Add v to buffer (cannot be directly added to the sample)

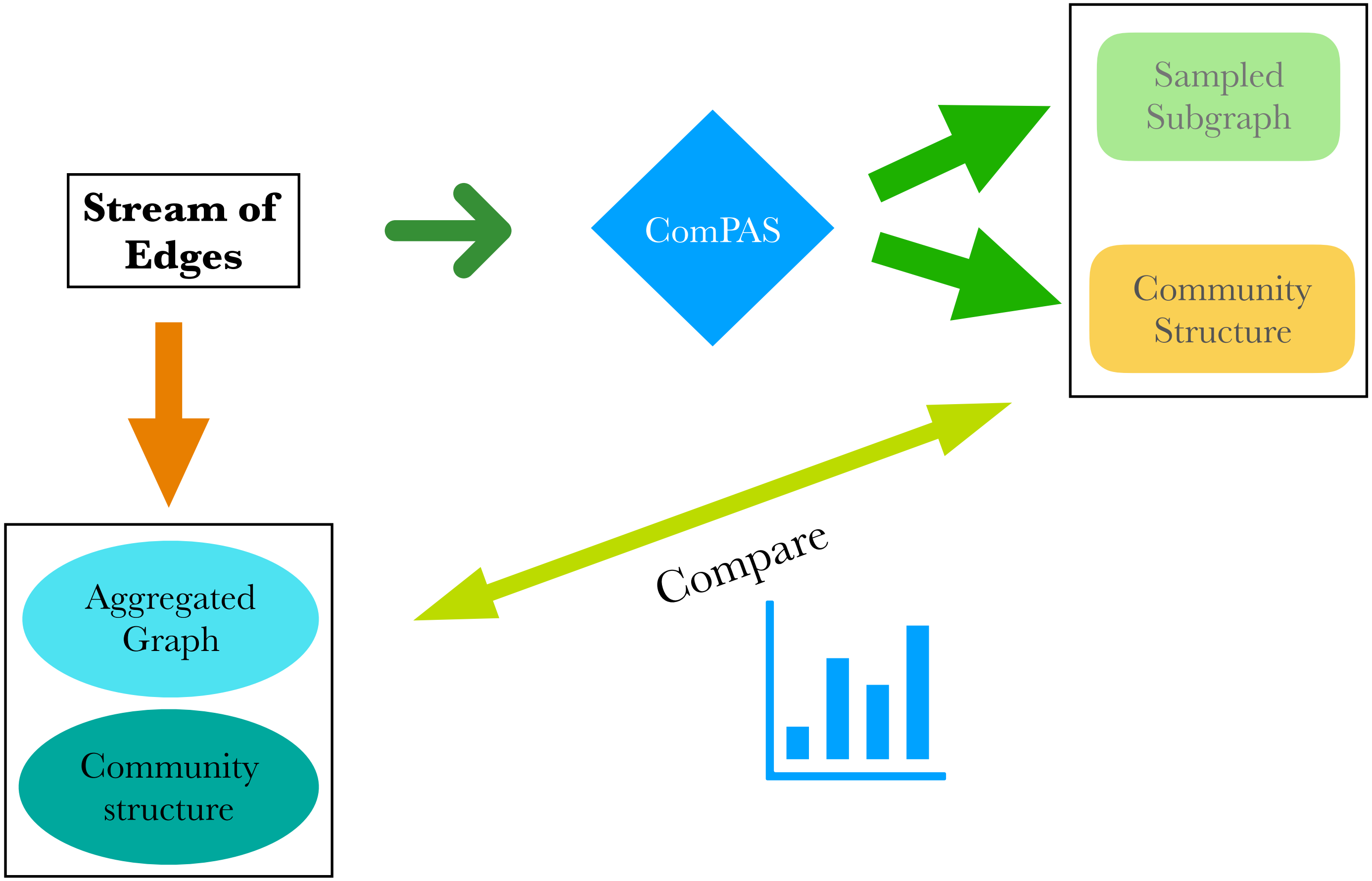
Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Subsequent cases

edge: (u,v)

- u is in sample and v is new
    - v is inserted into buffer which might trigger rearrangement of the buffer and sample
- u is in buffer and v is new
    - Increase $H_c(u)$ by 1
    - Insert v into buffer
- Both u and v are new
    - Insert both u and v into buffer

What do we have?



Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

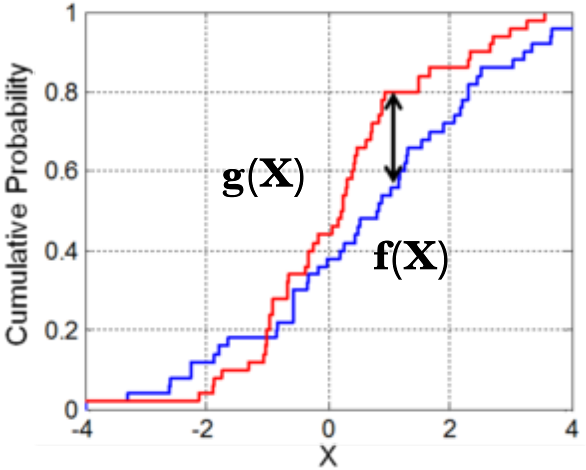
Evaluation

- Experiments performed on 4 real-world and 1 synthetic datasets
- Two ways of evaluation
  - Quality of the community structure
  - Content of the communities
- Baselines -
  - Streaming node (SN), streaming edge (SE), streaming BFS (SBFS) and Partially induced edge sampling (PIES)
  - Novel Green Algorithm (sample obtained on aggregated graph)

Evaluation

- Quality of community structure
  - Based on 13 topological measures proposed by Yang and Leskovec
  - Structural properties like average degree, internal density … (calculated for each community)

- We compare using D-statistics -
  - Consider a property X
  - Calculate distribution of X across communities in the ground-truth (f(X)) and the obtained sample g(X)
  - Calculate D-statistics between f(X) and g(X)



Evaluation

- Content of the community structure
- Similarity measured through -
  - Purity
  - Normalized Mutual Information (NMI)
  - Adjusted Rand Index (ARI)

ComPAS outperforms all other streaming graph sampling algorithm

Sikdar et. al, ComPAS: Community Preserving Sampling for Streaming Graphs. AAMAS 2018

Future directions

- Theoretical guarantees on the quality of the sample
- Complexity of the algorithm
- Allow deletion of edges over time

Thank You

Contact: Sandipan Sikdar
Email: sandipan.sikdar@cssh.rwth-aachen.de

Ref: Sandipan Sikdar, Tanmoy Chakraborty, Soumya Sarkar, Niloy Ganguly, Animesh Mukherjee: **ComPAS: Community Preserving Sampling for Streaming Graphs.** AAMAS 2018