A space efficient streaming algorithm for triangle counting using the birthday paradox

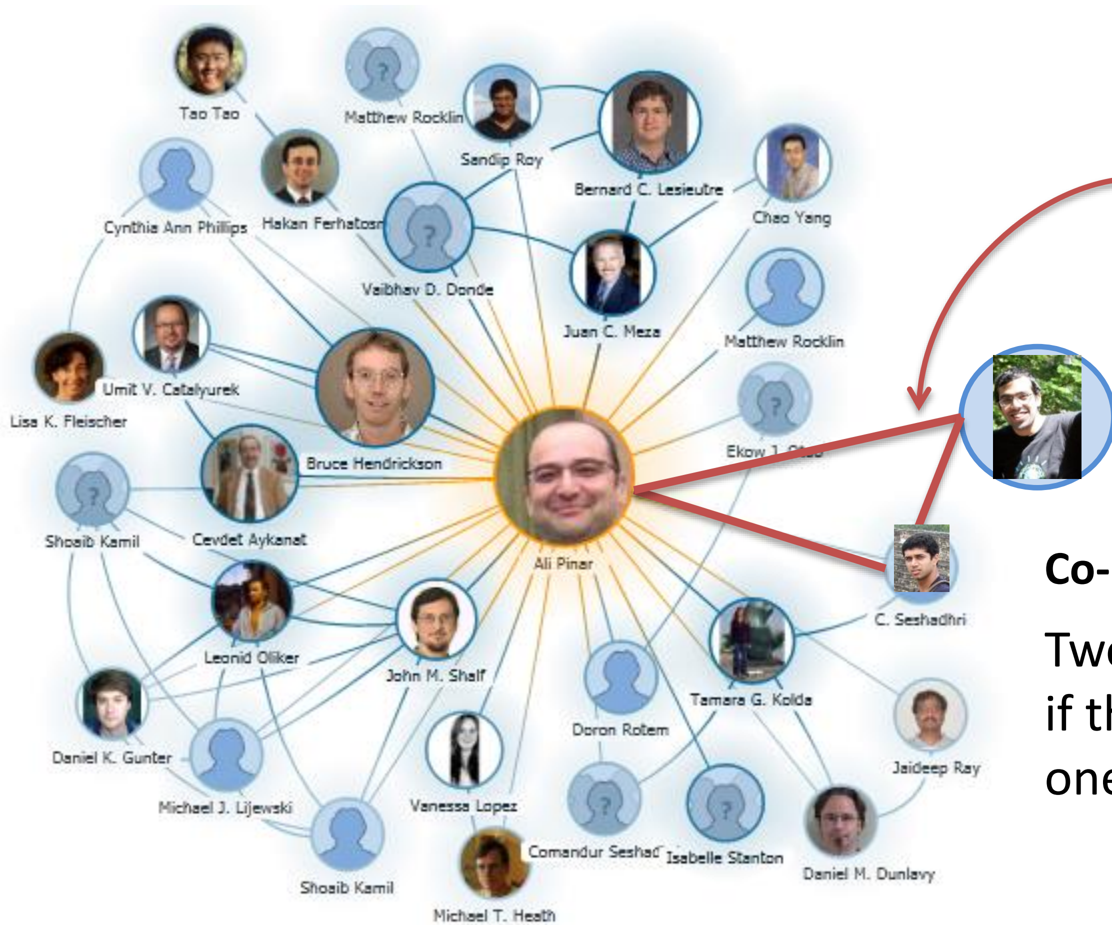
Madhav Jha

(Penn State → Sandia National Labs)

Joint work with **C. Seshadhri** (Sandia National Labs)
and **Ali Pinar** (Sandia National Labs)

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000."

Real-world graphs: An Example



This work!

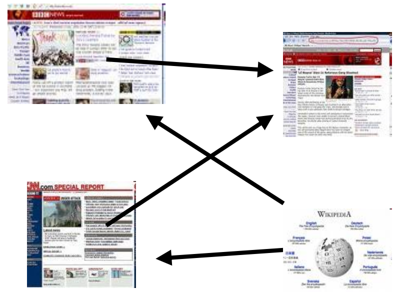
Co-authorship network

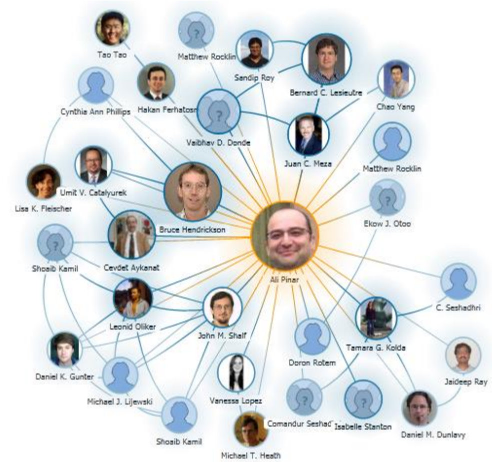
Two authors are connected if they publish at least one paper together

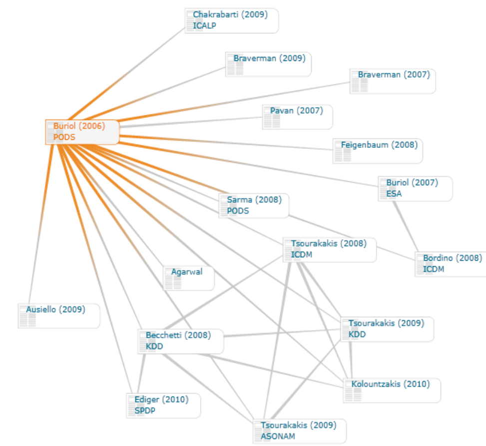
Graph [SNAP]	# nodes (n)	# edges (m)	# triangles (T)
Ca-HepPh	12K	118K	3.35M

Source: http://academic.research.microsoft.com

Real-world graphs





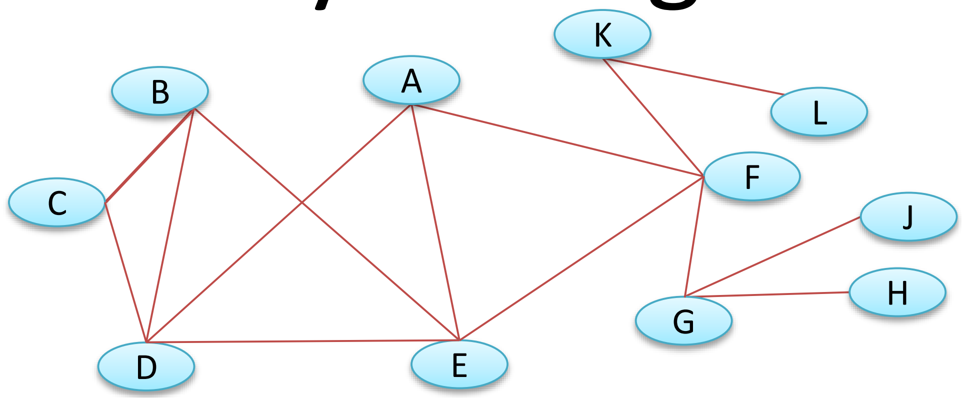


1. Graphs are everywhere.
2. Real-world graphs are huge. (Lots of vertices and edges.)
3. Real-world graphs have lots of triangles.

Graph [SNAP]	# nodes (n)	# edges (m)	# triangles (T)
web-BerkStan	0.6M	6M	64M
orkut	3M	22M	627M
Ca-HepPH	12K	118K	3.35M
cit-Patents	3M	16M	7M

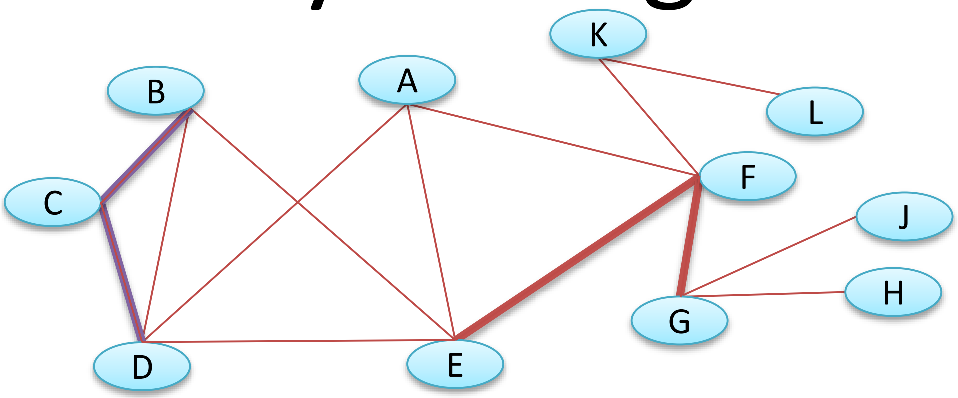
Photo Credit: a) facebook.com
b) http://academic.research.microsoft.com

Transitivity: Triangle "density"



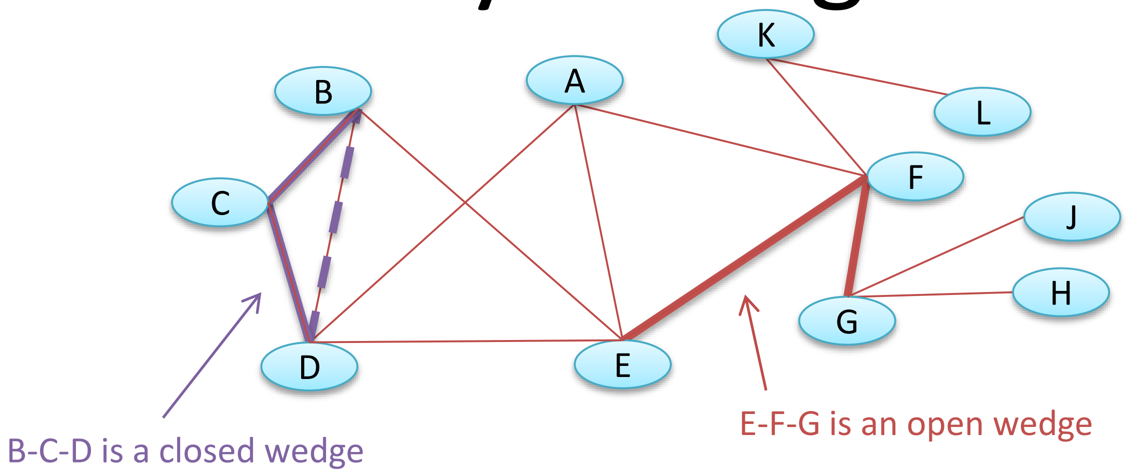
- A wedge is a length 2 path. Namely, a "potential" triangle.
- Transitivity = τ = 3 #Triangles/ #Wedges = fraction of closed wedges

Transitivity: Triangle "density"



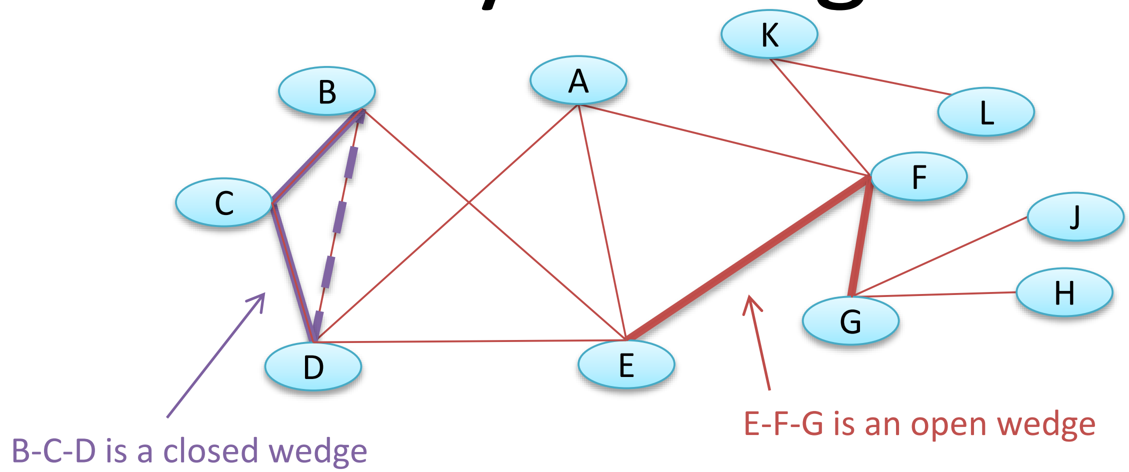
- A wedge is a length 2 path. Namely, a "potential" triangle.
- Transitivity = τ = 3 #Triangles/ #Wedges = fraction of closed wedges

Transitivity: Triangle “density”



- A wedge is a length 2 path. Namely, a “potential” triangle.
- Transitivity = τ = 3 #Triangles/ #Wedges = fraction of closed wedges

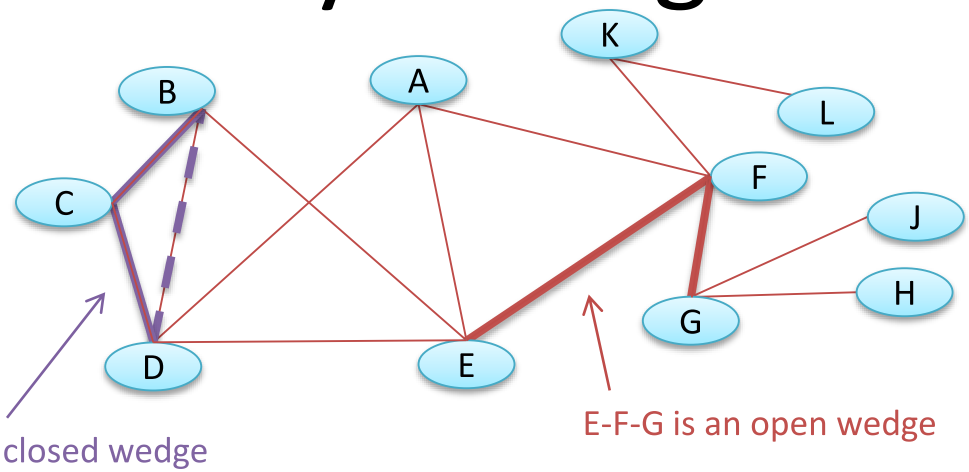
Transitivity: Triangle “density”



- A wedge is a length 2 path. Namely, a “potential” triangle.
- Transitivity = τ = 3 #Triangles/ #Wedges = fraction of closed wedges

Graph [SNAP]	# nodes (n)	# edges (m)	# triangles (T)	Transitivity
web-BerkStan	0.6M	6M	64M	0.007
orkut	3M	223M	627M	0.041
Ca-HepPH	12K	118K	3.35M	0.39
cit-Patents	3M	16M	7M	0.067

Transitivity: Triangle "density"



- A wedge is a length 2 path. Namely, a "potential" triangle.
- Transitivity = τ = 3 #Triangles/ #Wedges = fraction of closed wedges

[Seshadhri Pinar Kolda 2013] gave algorithm for computing transitivity given accesss to the entire graph. This algorithm is the starting point of of work.

Graph [SNAP]	# nodes (n)	# edges (m)	# triangles (T)	Transitivity
web-BerkStan	0.6M	6M	64M	0.007
orkut	3M	223M	627M	0.041
Ca-HepPH	12K	118K	3.35M	0.39
cit-Patents	3M	16M	7M	0.067

Why Count Triangles in Graphs?

- Useful in Social Science for positing various theses on behavior [Burt 09], [Coleman 88], [Welles, Devender, Contractor 10], [Portes 88]
- Applied to spam detection [Becchetti Boldi Castillo Gionis 08]
- Relevant for finding topics on WWW [Eckmann Moses 02]
- Proposed as a guide for community structure
  Stated as a core feature for graph models [Vivar Banks 11]
  Cornerstone for Block Two-level Erdos-Renyi (BTER) [Seshadhri Pinar Kolda 12]
- Good descriptor of the underlying graph [Durak Seshadhri Pinar Kolda 12]
- Rich set of algorithmic results spanning various models
  (exact/approximate/deterministic/randomized/…) X (streaming, map-reduce, parallel etc.)
- Very well-studied: [Ahn Guha McGregorGraph 2012], [Durak Pinar Kolda Seshadhri 2012], [Pagh Tsourakakis 2012], [Suri Vassilvitskii 2011], [Tsourakakis Kolountzakis Miller 2011], [Chu Cheng 2011], [Yoon Kim 2011][Kolountzakis Miller Peng Tsourakakis 2010], [Avron 2010],[Tsourakakis Drineas Michelakis Koutis Faloutsos 2009], [Tsourakakis Kang Miller Faloutsos 2009], [Latapy 2008], [Becchetti Boldi Castillo Gionis 2008], [Tsourakakis 08], [Buriol Frahling Leonardi Marchetti-Spaccamela Sohler 2006], [Jowhari Ghodsi 2005], [Schank Wagner 2005], [Bar-Yossef Kumar Sivakumar 2002], …

Graph as stream of edges

- Real-world graphs have a natural time-stamp

Madhav Jha and Saravanan Va are now friends.
4 August at 02:01

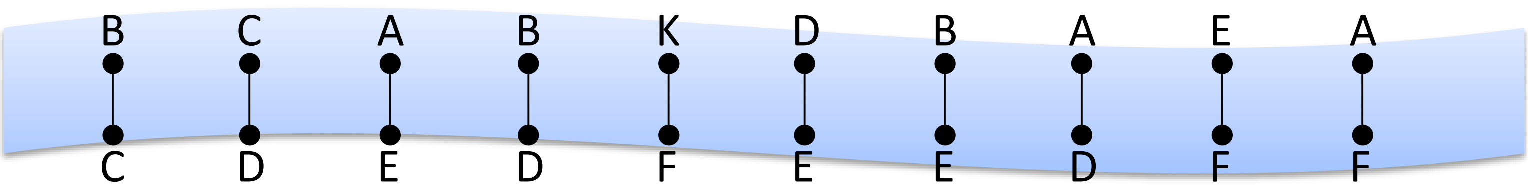
Madhav Jha and Ratul Ray are now friends.
21 July at 01:54

⋮

Madhav Jha and Hemant Kumar are now friends.
7 July at 00:11

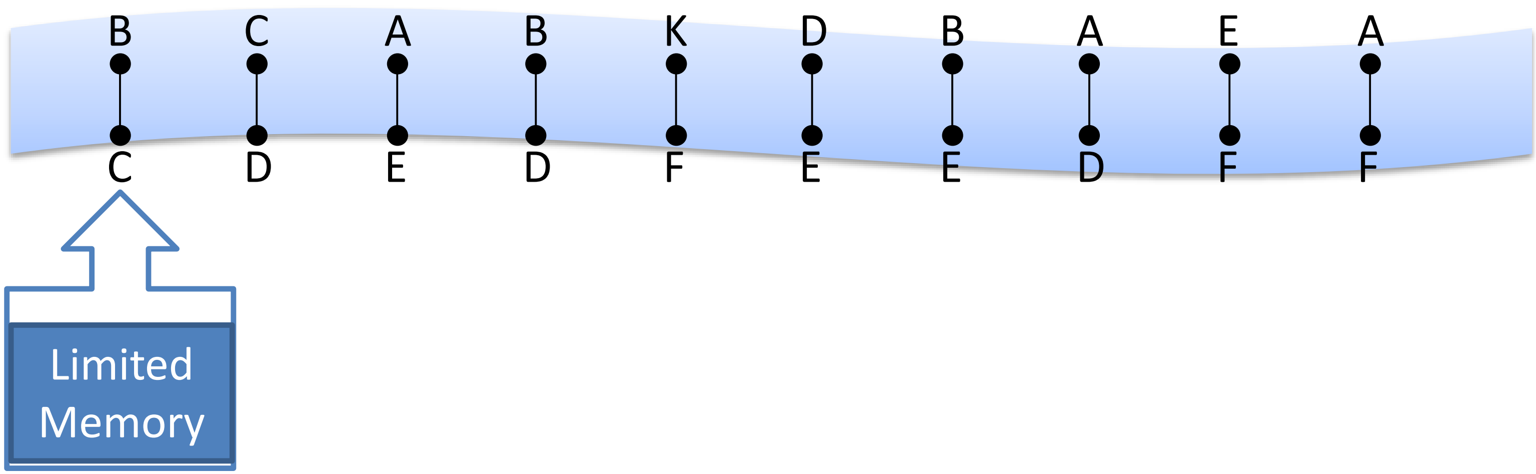
Photo credit: facebook.com

Graph as stream of edges



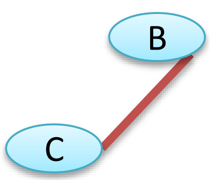
Triangles so far:
Graph seen so far:

Graph as stream of edges

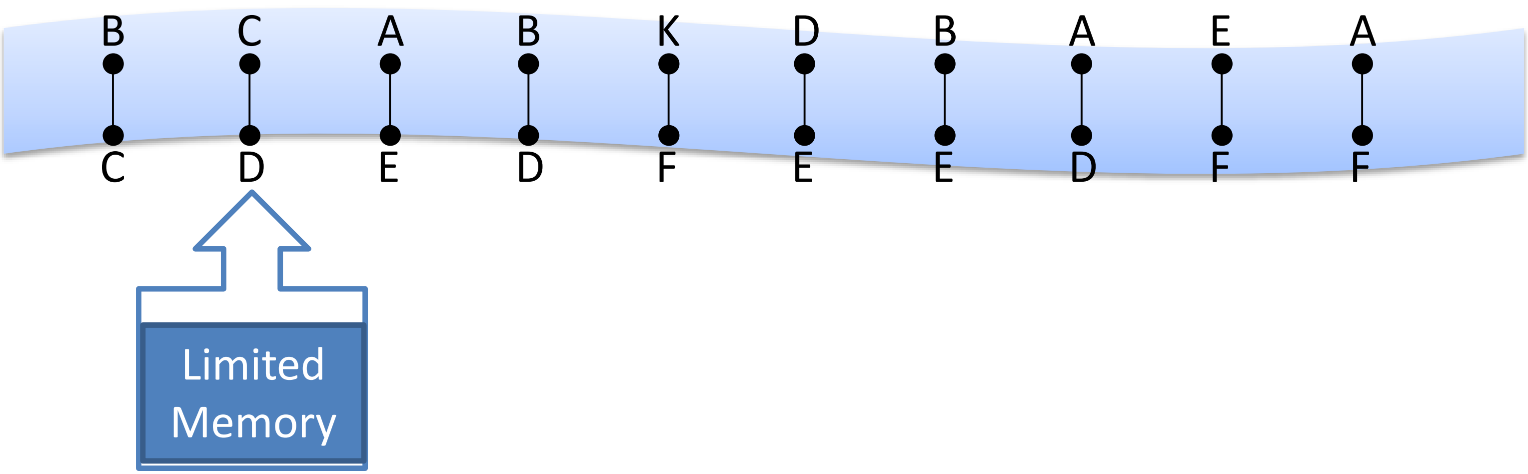


Triangles so far:

Graph seen so far:

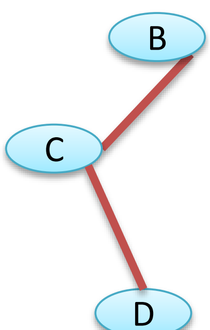


Graph as stream of edges

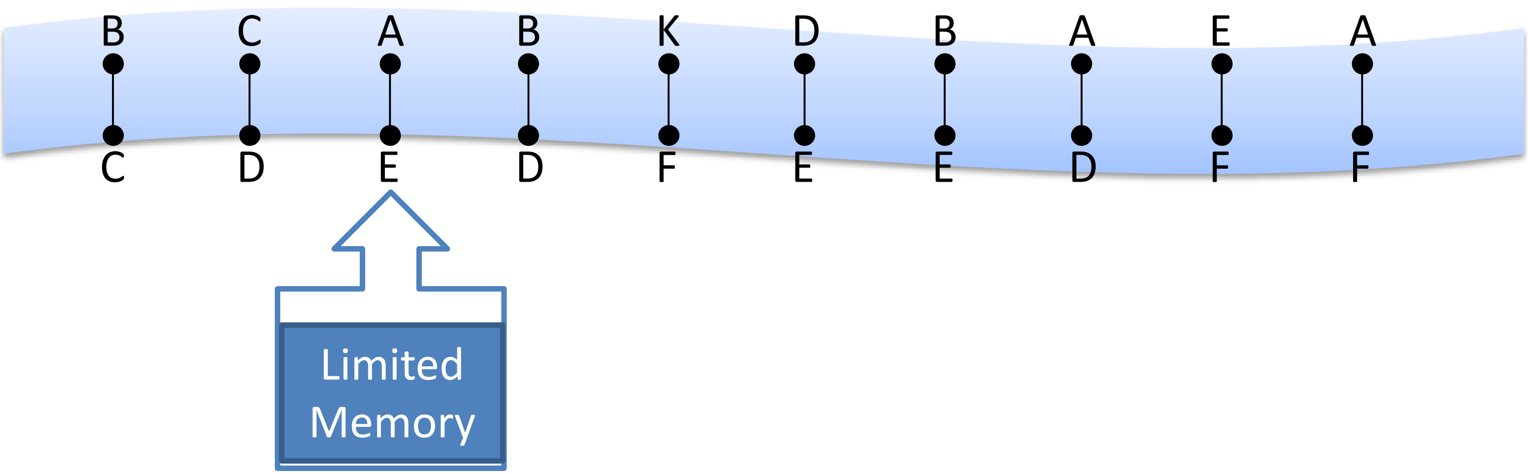


Triangles so far:

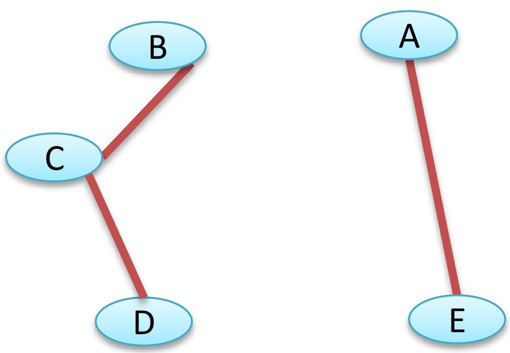
Graph seen so far:



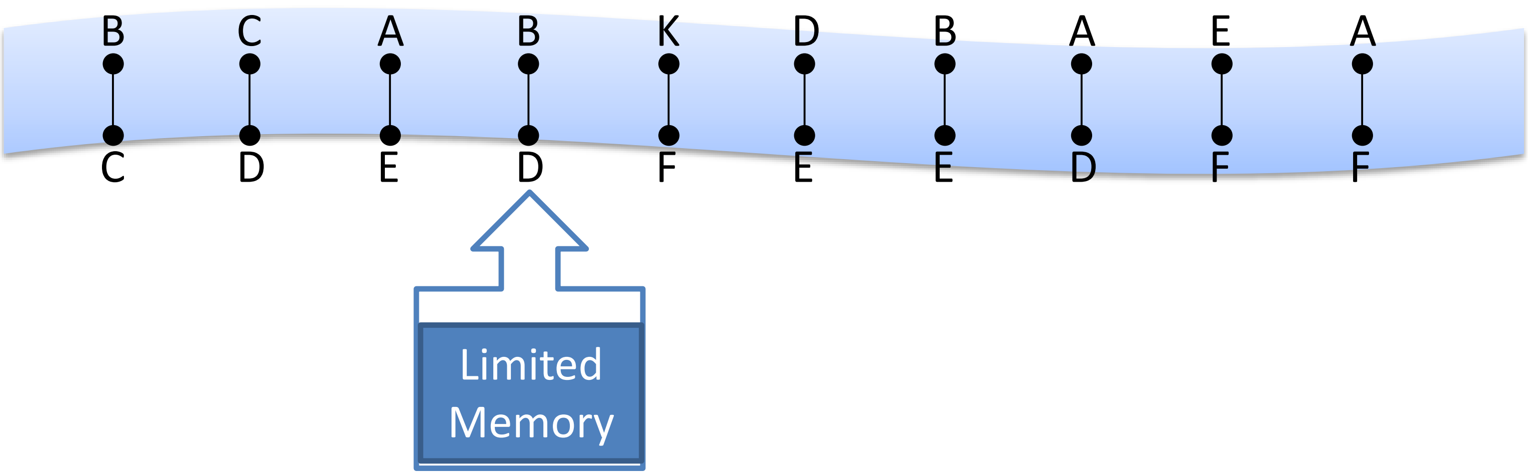
Graph as stream of edges



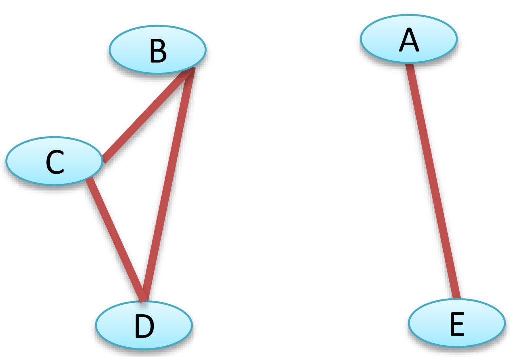
Triangles so far:
Graph seen so far:



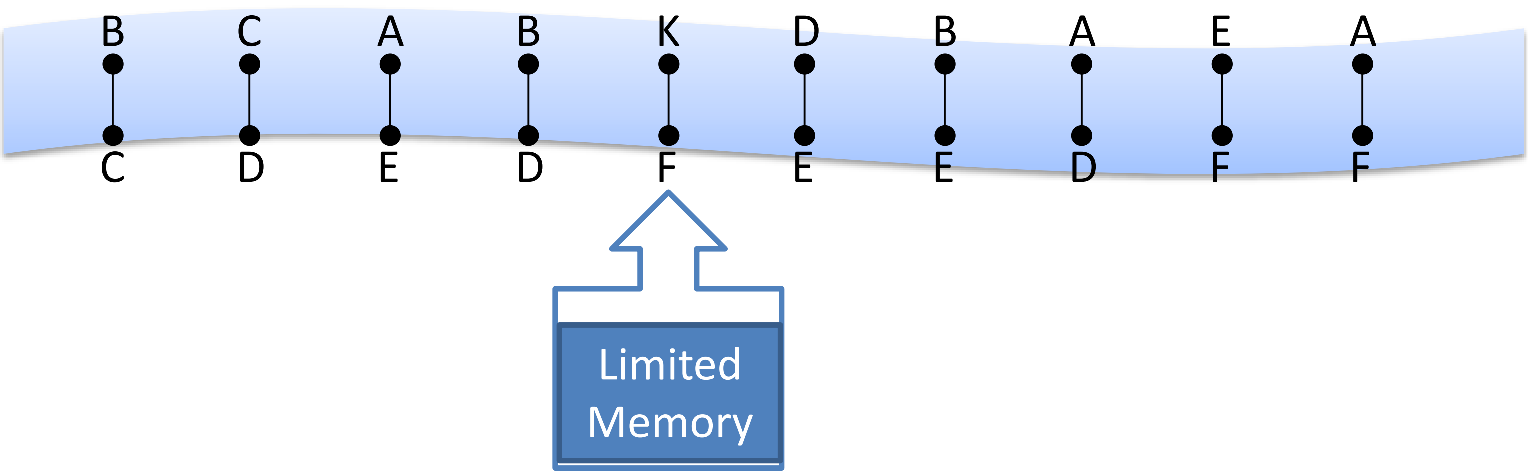
Graph as stream of edges



Triangles so far: 1
Graph seen so far:

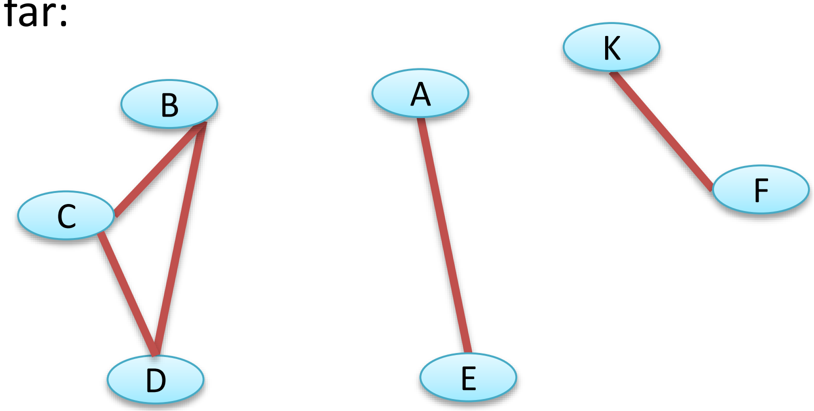


Graph as stream of edges

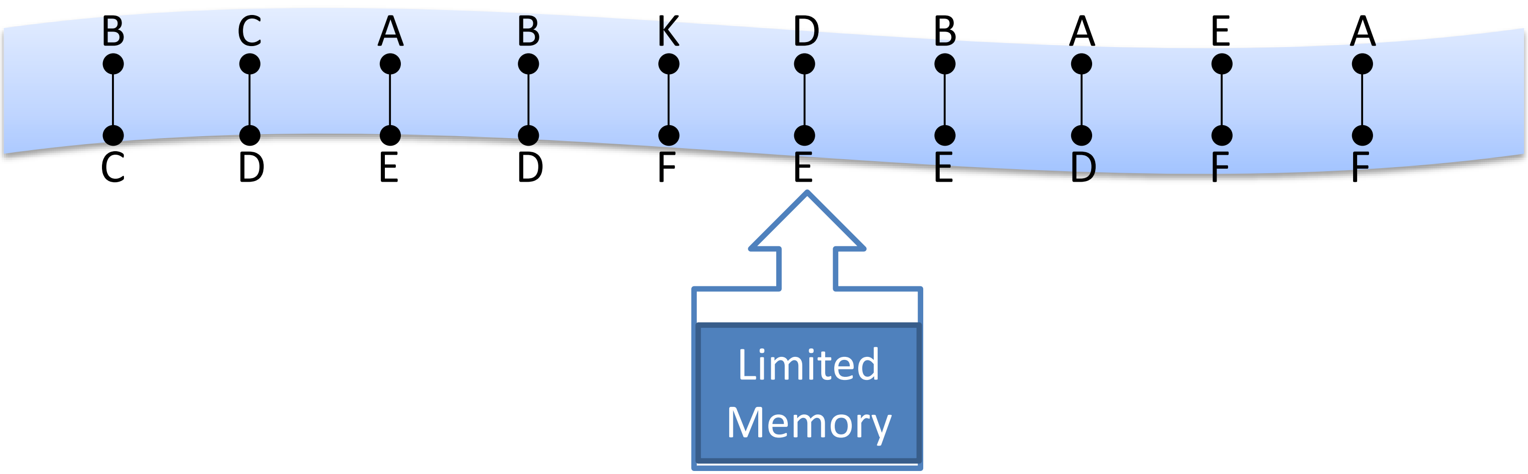


Triangles so far: 1

Graph seen so far:

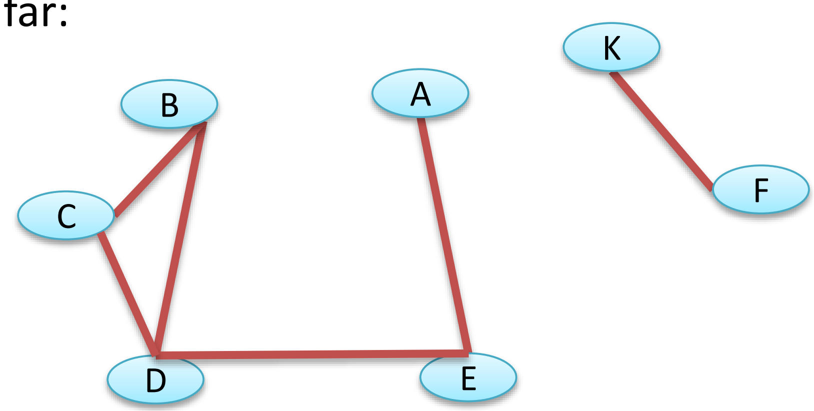


Graph as stream of edges

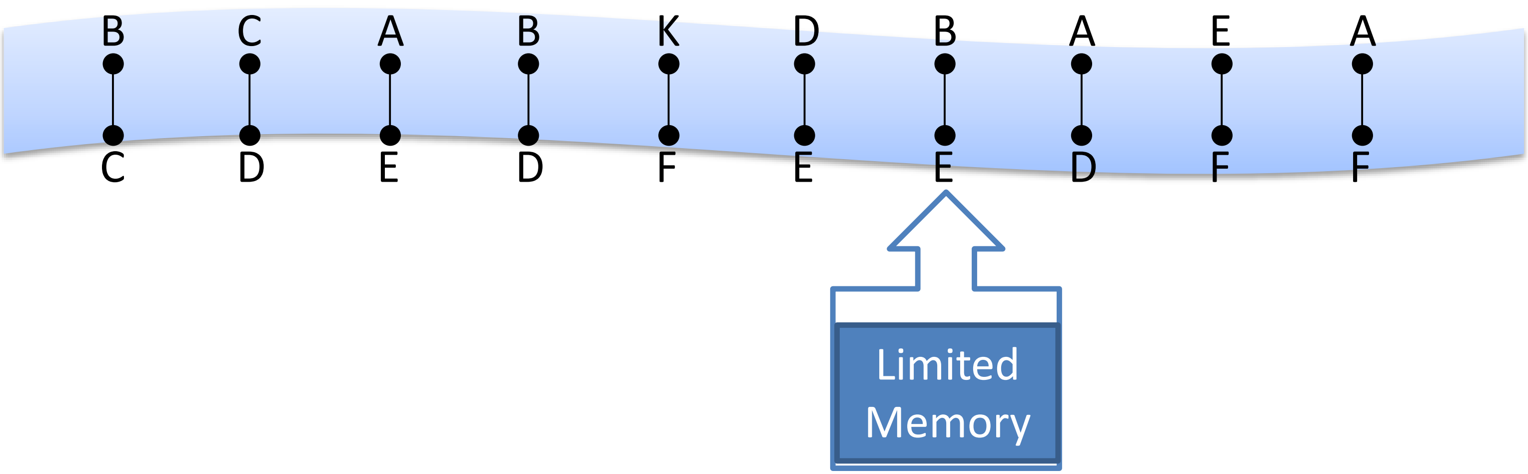


Triangles so far: 1

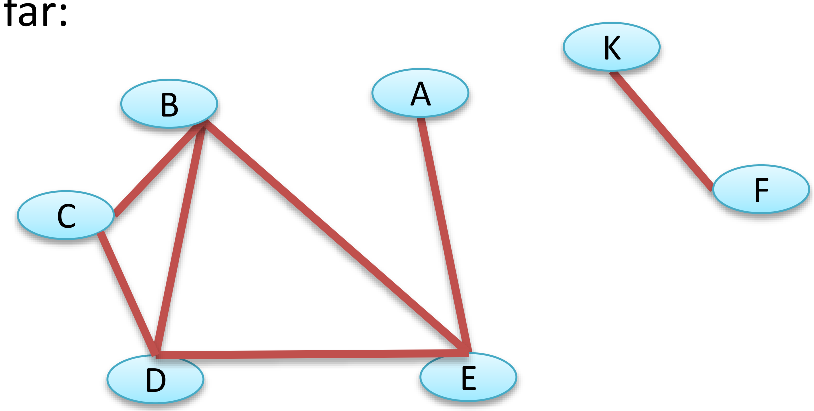
Graph seen so far:



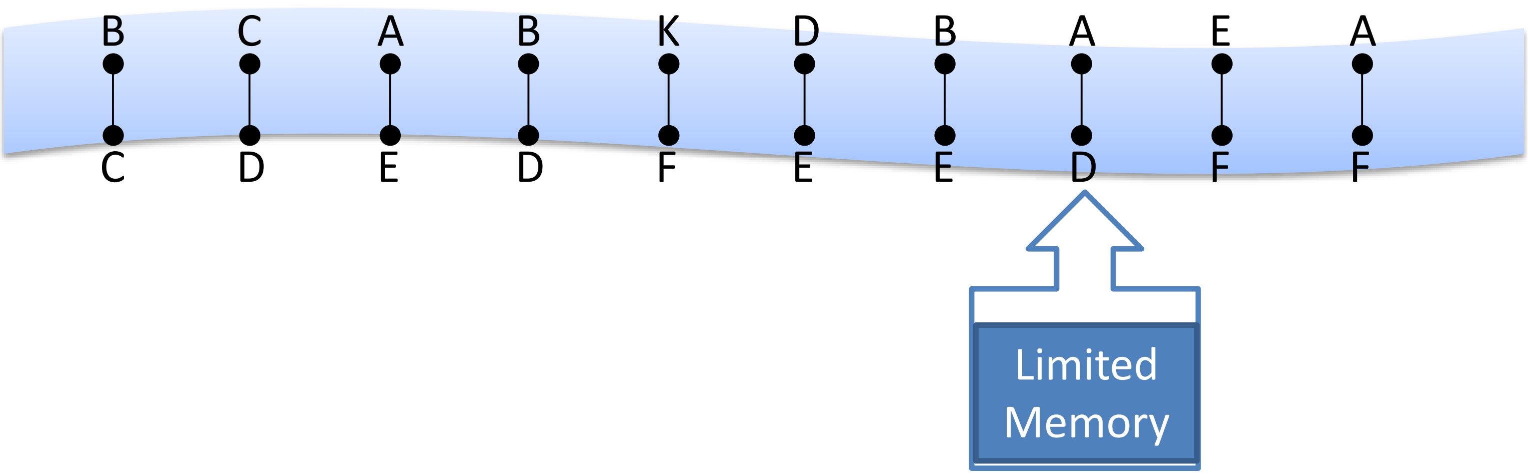
Graph as stream of edges



Triangles so far: 2
Graph seen so far:

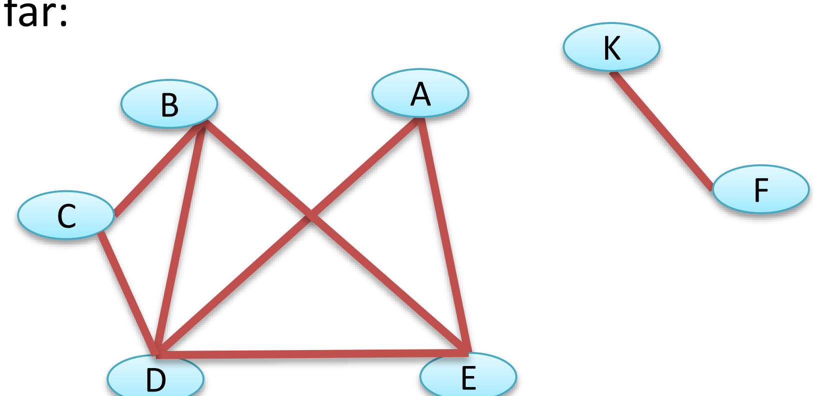


Graph as stream of edges

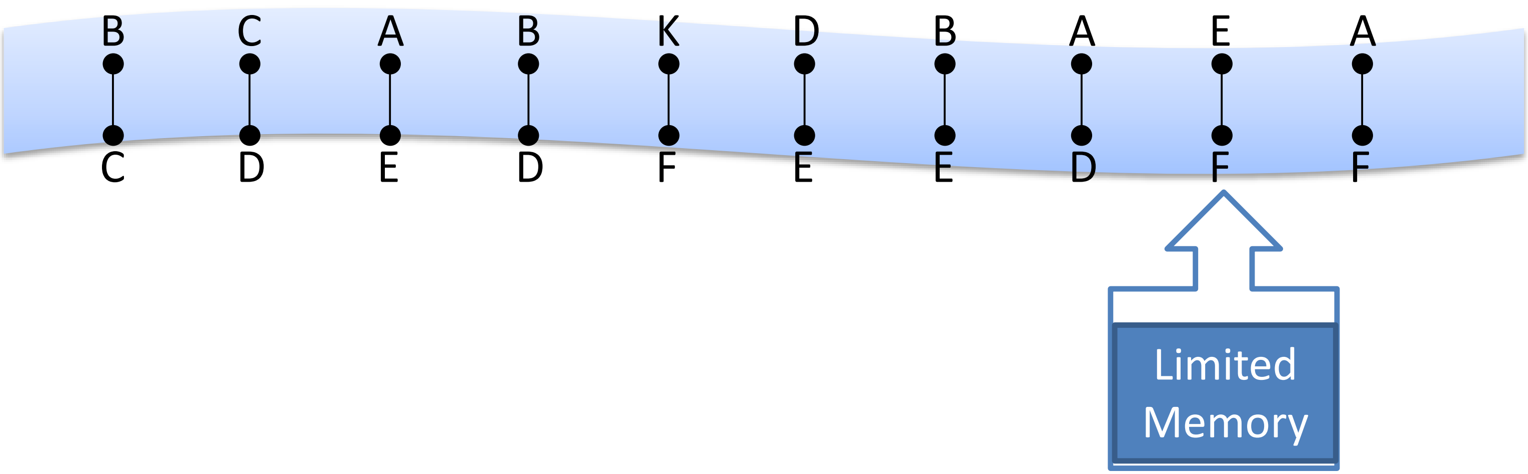


Triangles so far: 3

Graph seen so far:

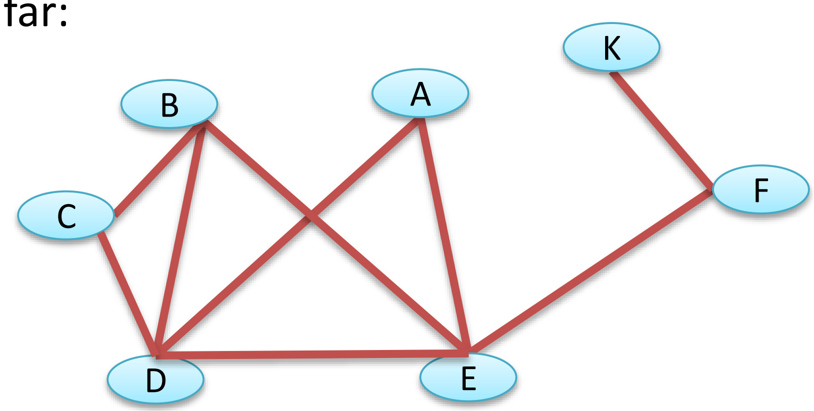


Graph as stream of edges

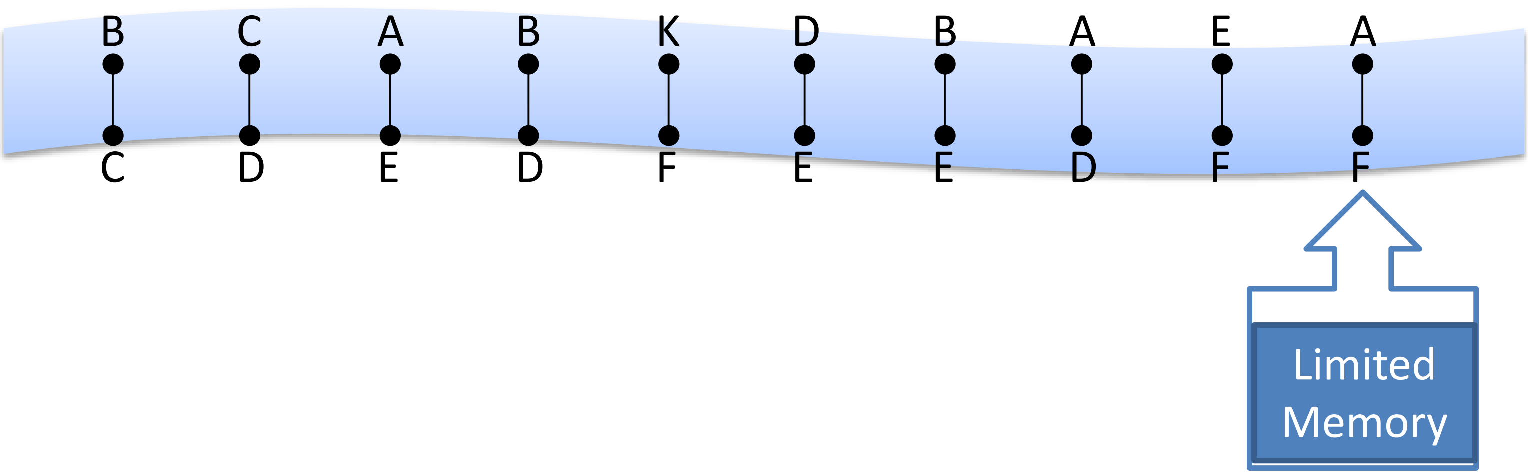


Triangles so far: 3

Graph seen so far:

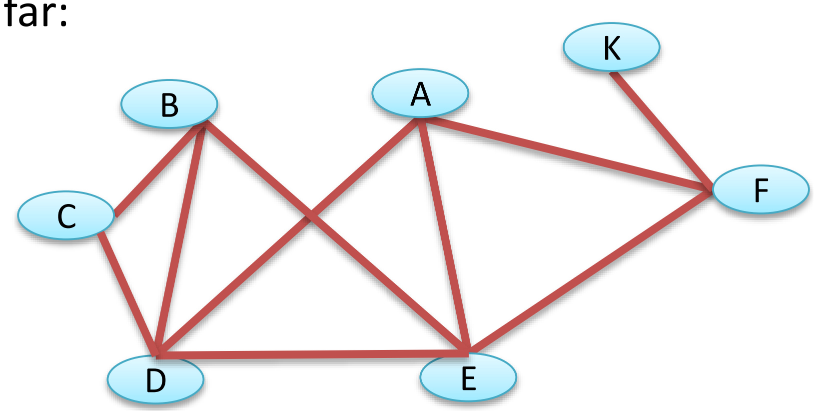


Graph as stream of edges



Triangles so far: 4

Graph seen so far:



Our Contributions : Theoretical

Theorem:

A single-pass streaming algorithm (for arbitrarily ordered edge stream) which stores only $O(\sqrt{n})$ edges (for most real world graphs), requires nearly constant time update per edge, and estimates # triangles and transitivity.

Analysis based on the classic Birthday Paradox.

Our Contributions : Practical

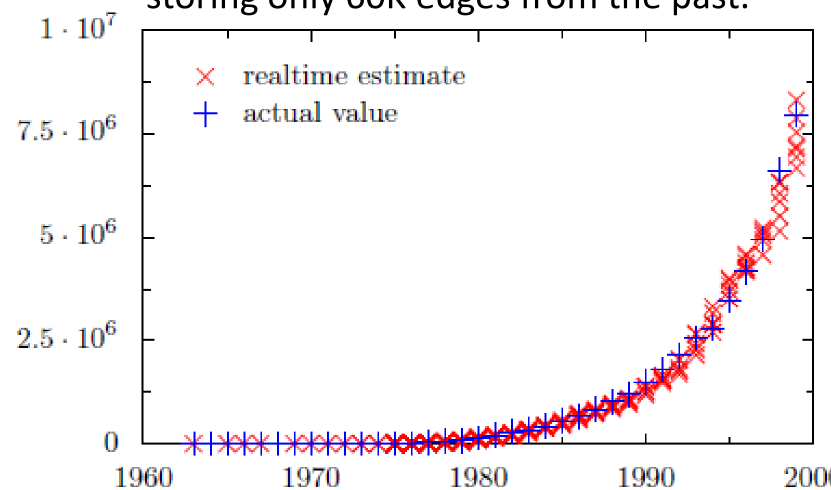
- Accurate triangles estimates in low space

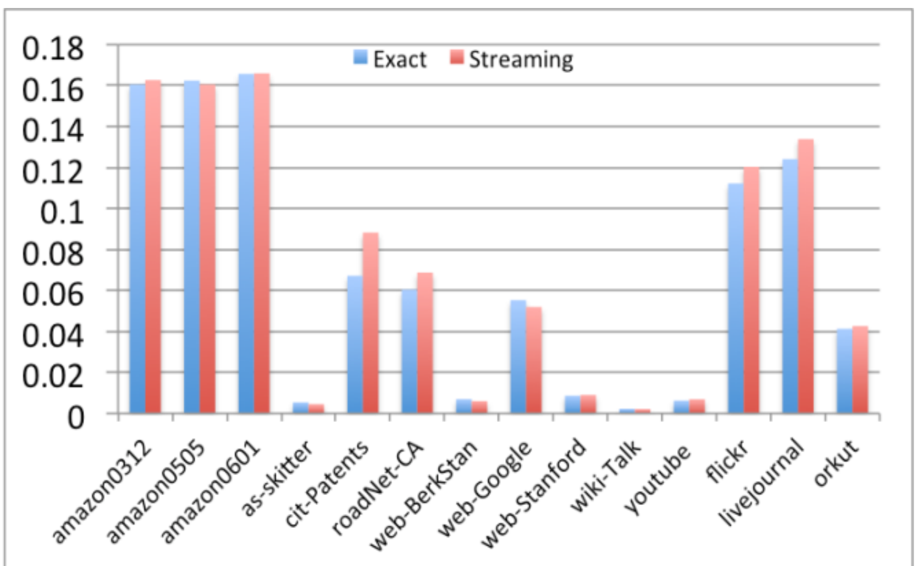
  Example: On Orkut graph (200 M edges and 0.627 B triangles), our algorithm stores only 40 K edges (2% of graph) and reports 0.658 B triangles (less than 5% relative error).

- Accurate transitivity estimates

- Realtime tracking

Realtime tracking of # triangles on cit-Patents graph (16M edges), storing only 60K edges from the past.



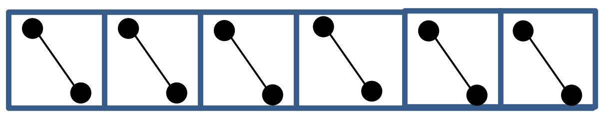


Estimating transitivity on a variety of dataset.
(Our algorithm stores only 40 K edges in all these runs.)

Data Structures of the Algorithm

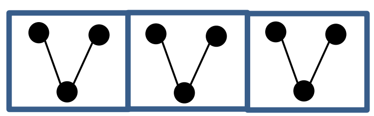
Input Parameters: s_e and s_w.

edge_reservoir[]



An array to store edges of size s_e

wedge_reservoir[]



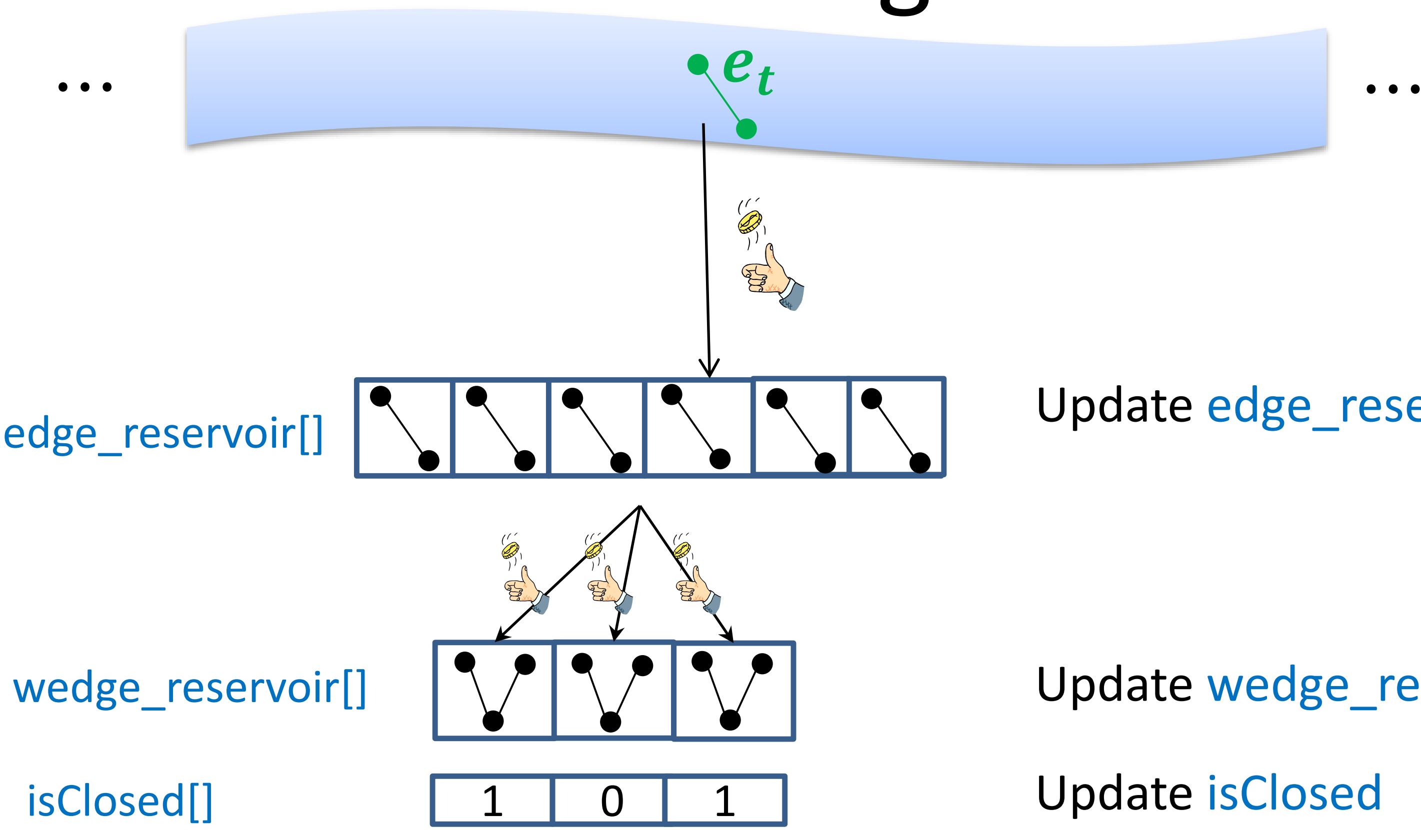
An array to store wedges of size s_w

isClosed[]

1	0	1

A Boolean array of size s_w

The Algorithm



Update edge_reservoir

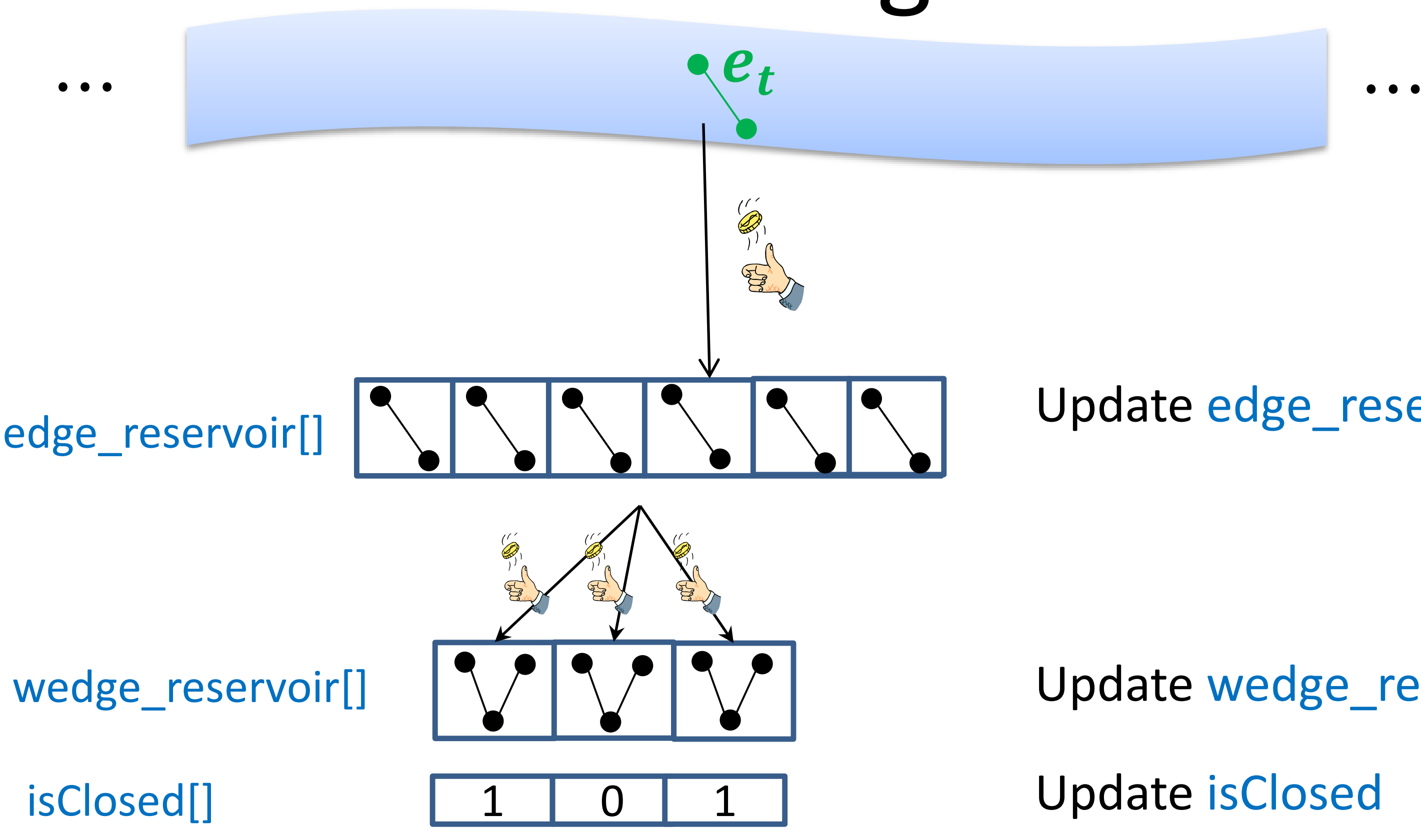
Update wedge_reservoir

Update isClosed

Let p be fraction of 1's in isClosed[]. Output

1. Transitivity, est-$\tau_t = 3p$
2. Triangles, est-$T_t =$ est-$\tau_t \times$ normalizing-factor

The Algorithm



Update edge_reservoir

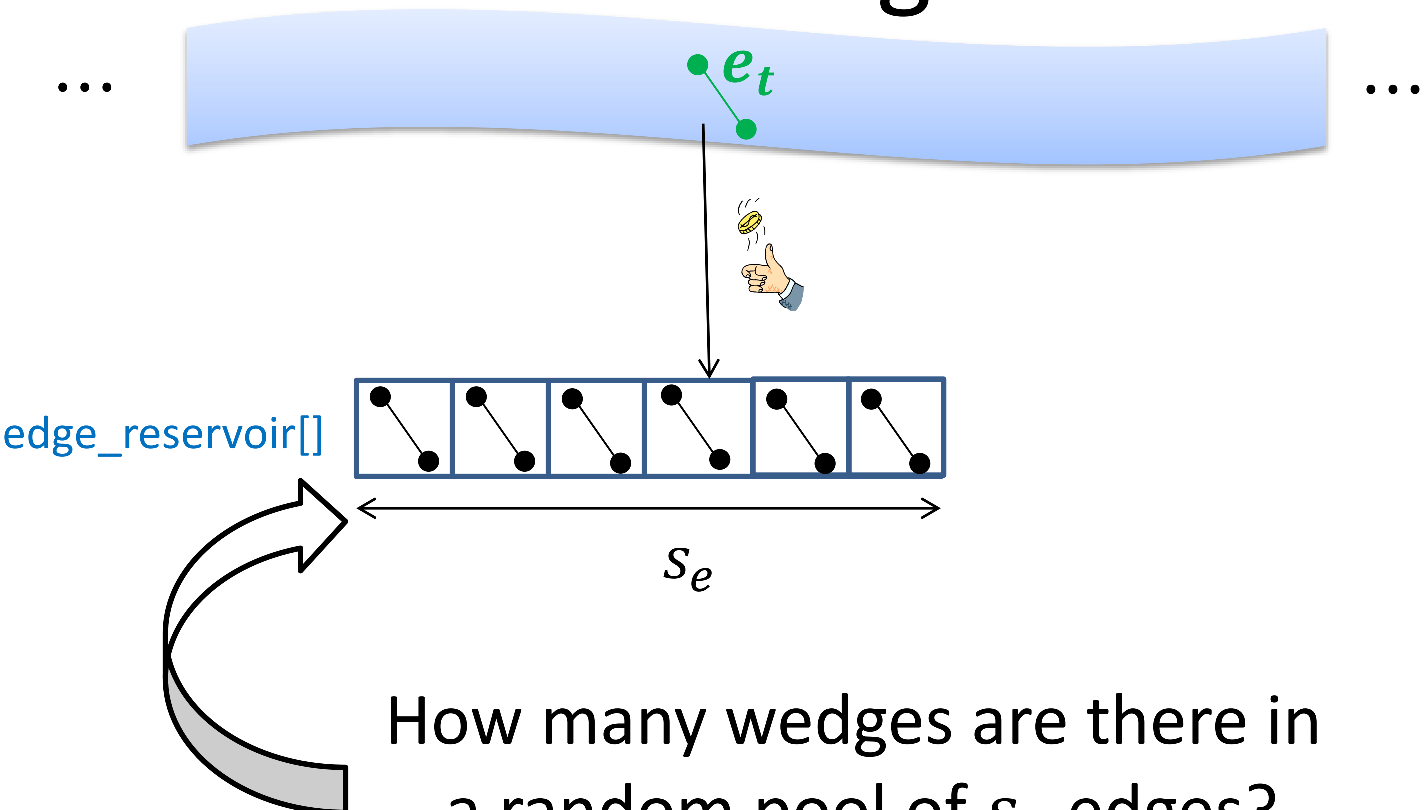
Update wedge_reservoir

Update isClosed

Updates to edge_reservoir very rare!

$$\sum_{t \leq m} 1 - (1 - 1/t)^{s_e} \approx \sum_{t \leq m} s_e/t \approx s_e \ln m$$

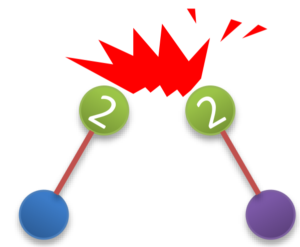
The Algorithm



How many wedges are there in a random pool of s_e edges?

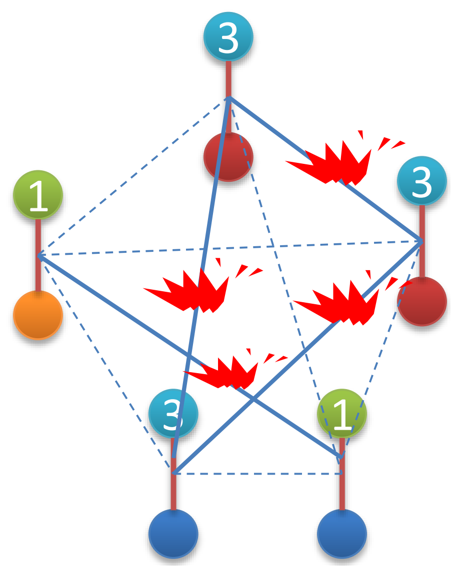
The Birthday Paradox to Rescue

Idea: Fundamentally, a wedge is a collision of two edges!



Birthday Paradox ⇒ s_e edges give rise to

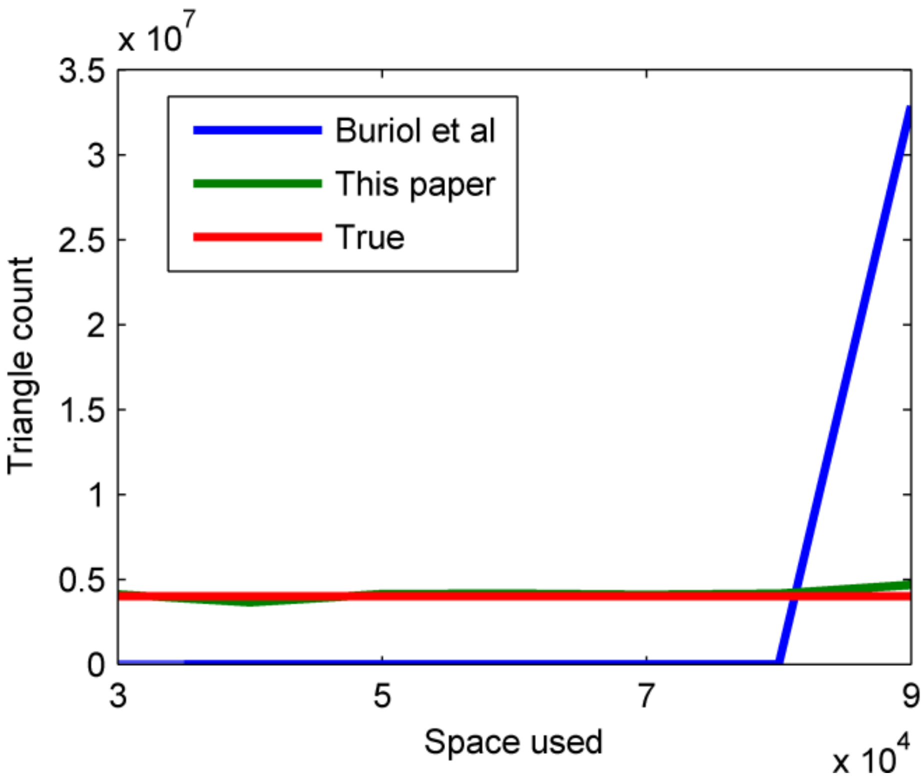
$$s_e^2 \cdot \Pr[\text{A single collision}]$$



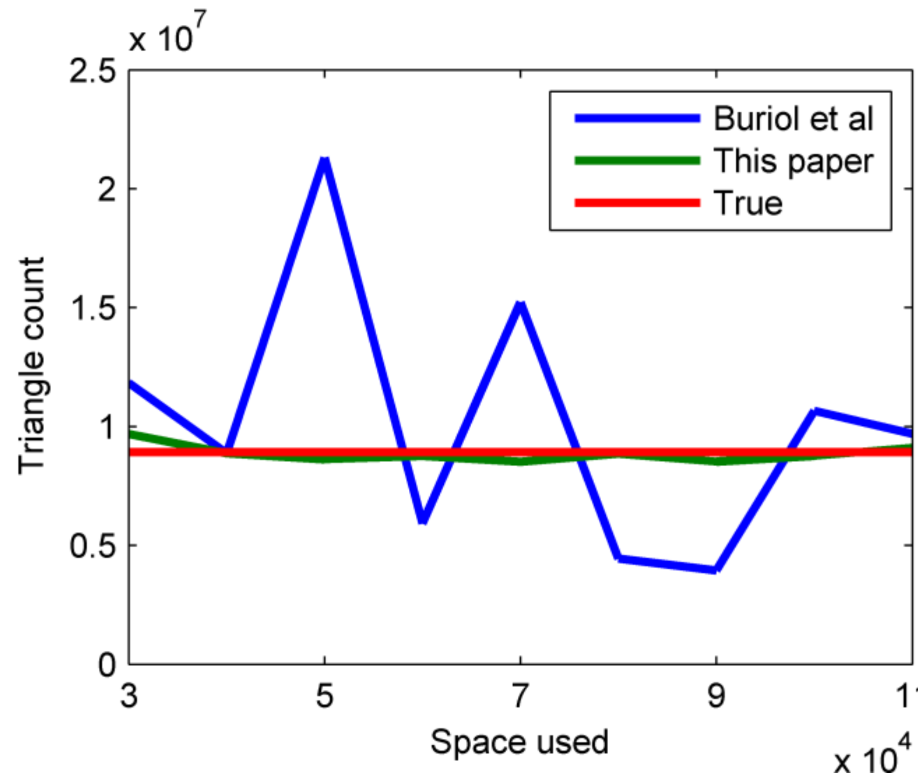
Experimental Results

Our Algorithm vs Buriol et al

Dataset: web-NotreDame



Dataset: amazon0505

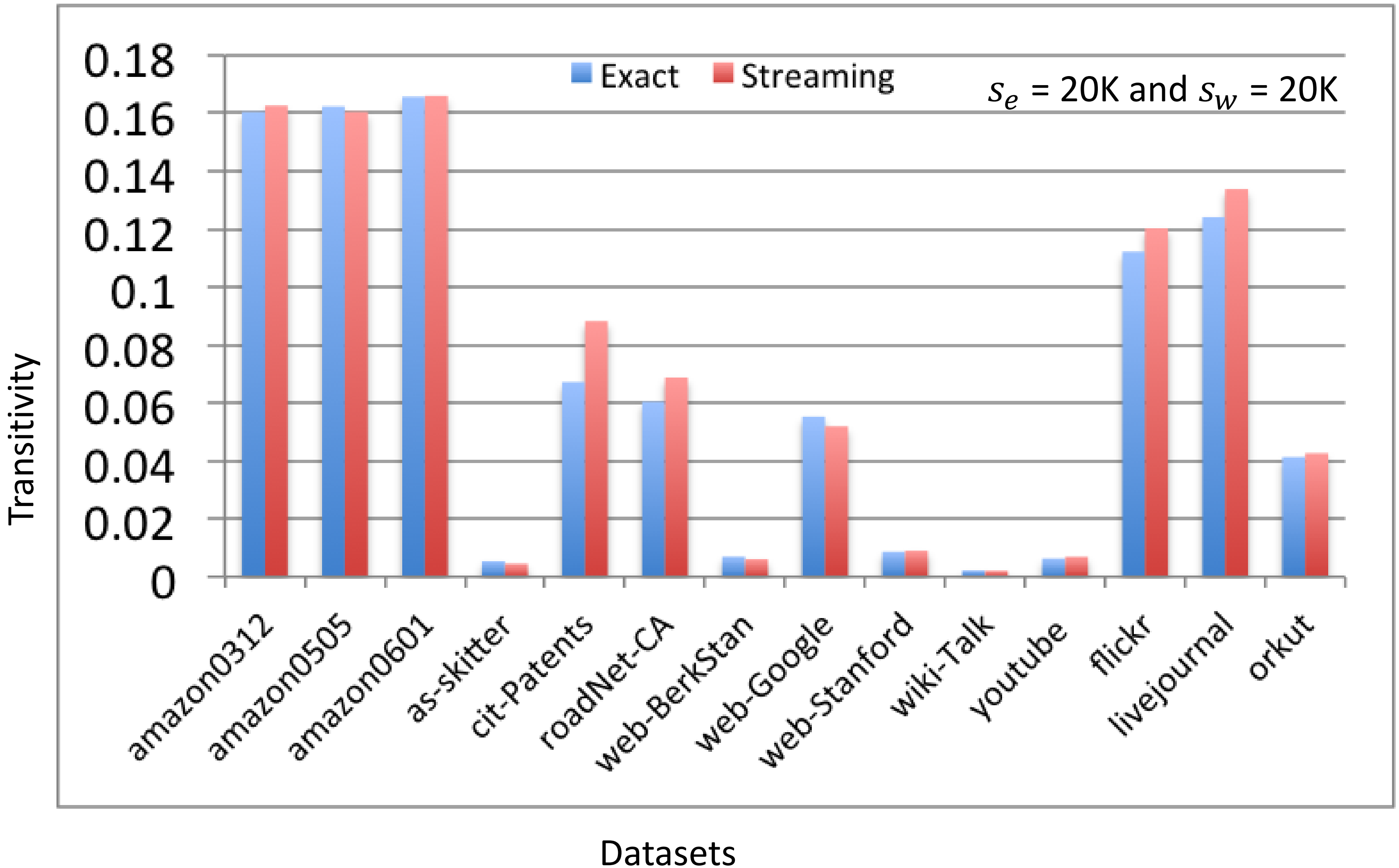


We fix s_e = 20K and vary s_w

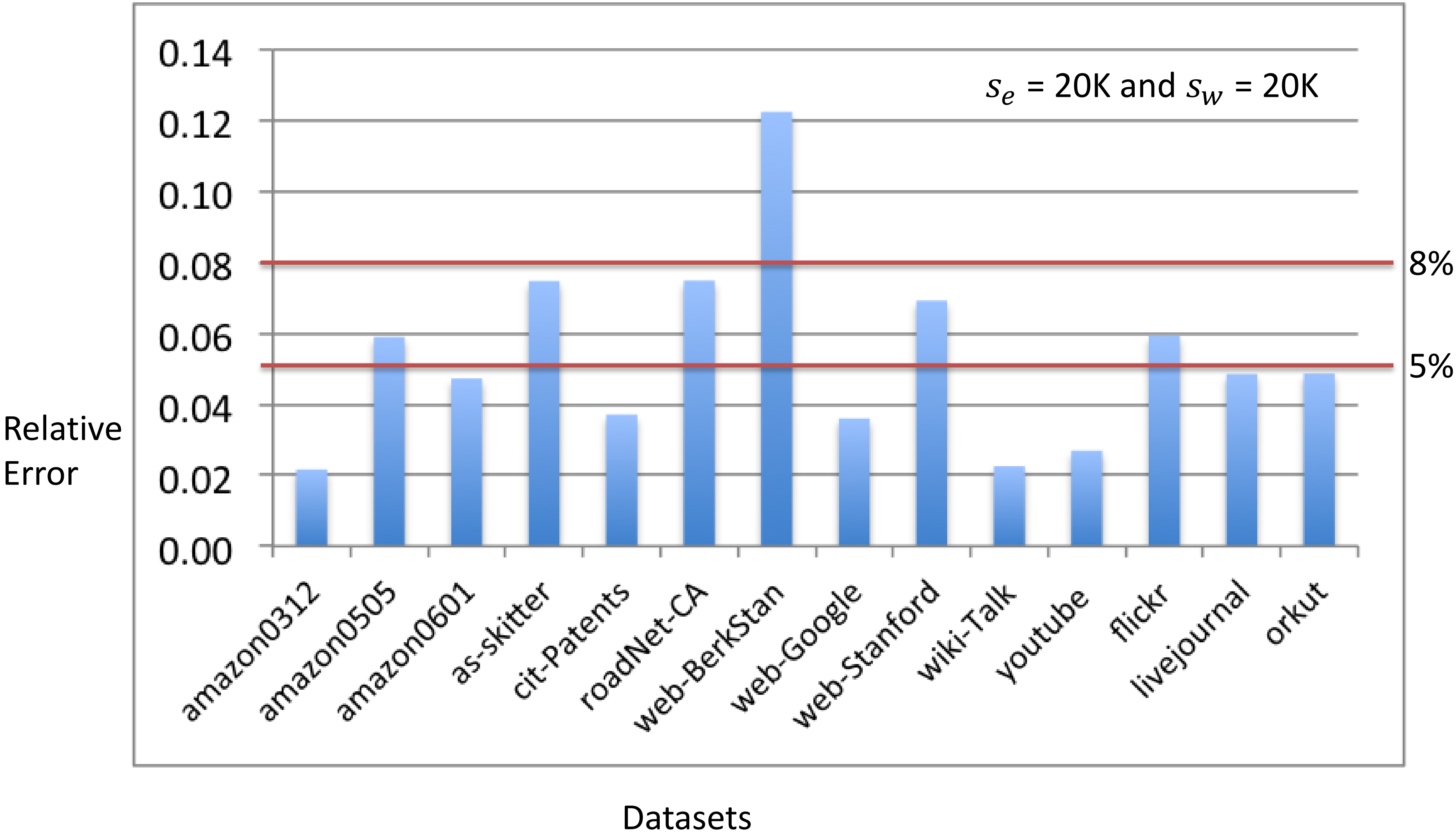
Space used in our algorithm: $s_e + s_w$
Space used in Buriol et al: number of edges sampled

Note: The results for Buriol et al is consistent with the analysis and experiments of their paper.

Accuracy of Transitivity Estimate



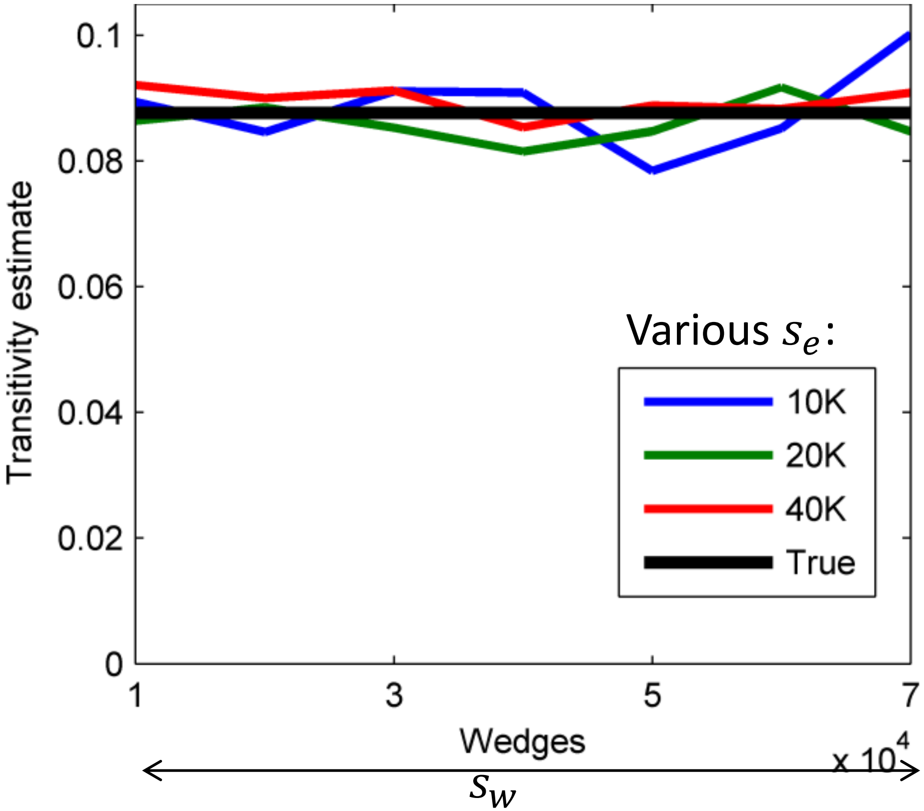
Accuracy of Triangles Estimate

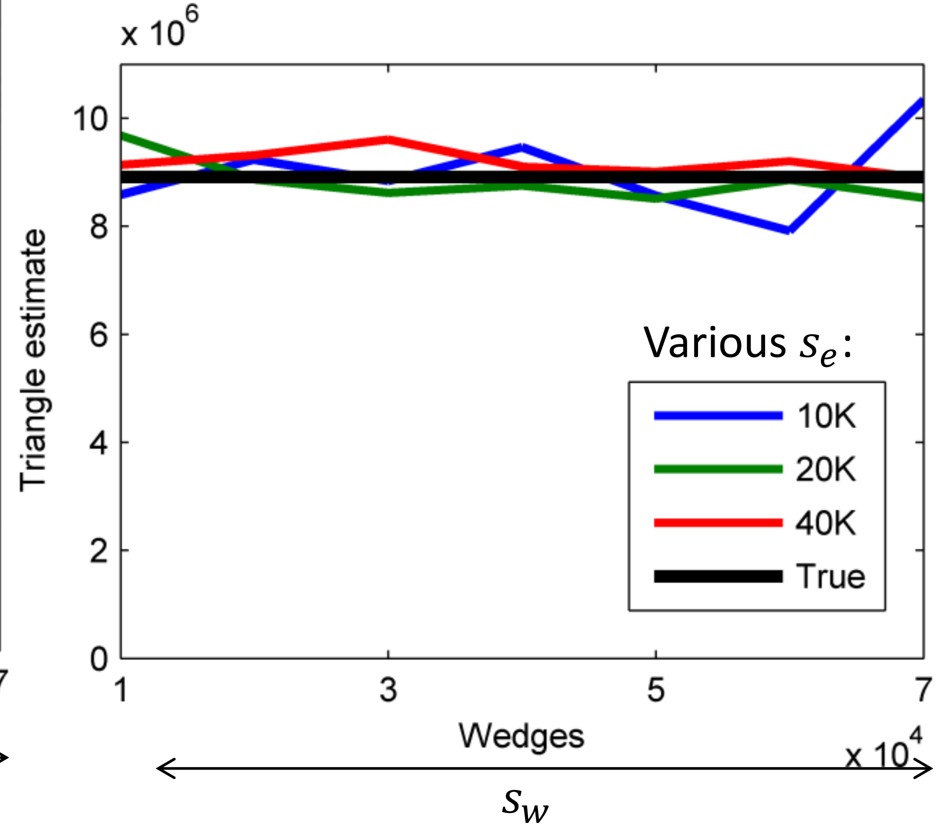


Note: web-BerkStan has very low transitivity 0.007. Therefore, relative error is high.

Convergence of Estimates

Dataset: amazon0505





Future Work

- Can we go below $\sqrt{n}$ space bound?
- Can we prove a lower bound on the space required by a 1-pass streaming algorithm to estimate triangle counts?
- Can we extend this approach to handle edge deletions ?
- Can we compute (and track) degree-wise clustering coefficient?