Triangle counting in dynamic graph streams

Konstantin Kutzkov and Rasmus Pagh

IT UNIVERSITY OF COPENHAGEN

Work supported by:

DET FRIE FORSKNINGSRÅD

erc

European Research Council

Established by the European Commission

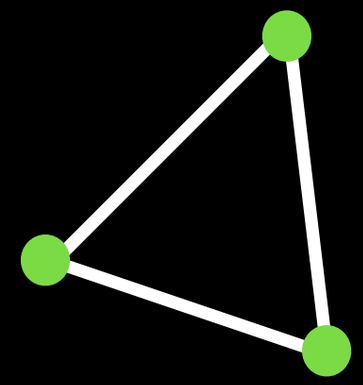
Supporting top researchers from anywhere in the world

Agenda

- Problem description and known results.
- Sampling-based approaches:
  - 2-path sampling
  - Edge sampling (Doulion, colorful sampling)
- New algorithm

Triangle counting

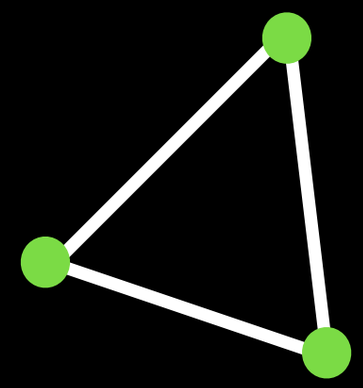
m edges, n nodes



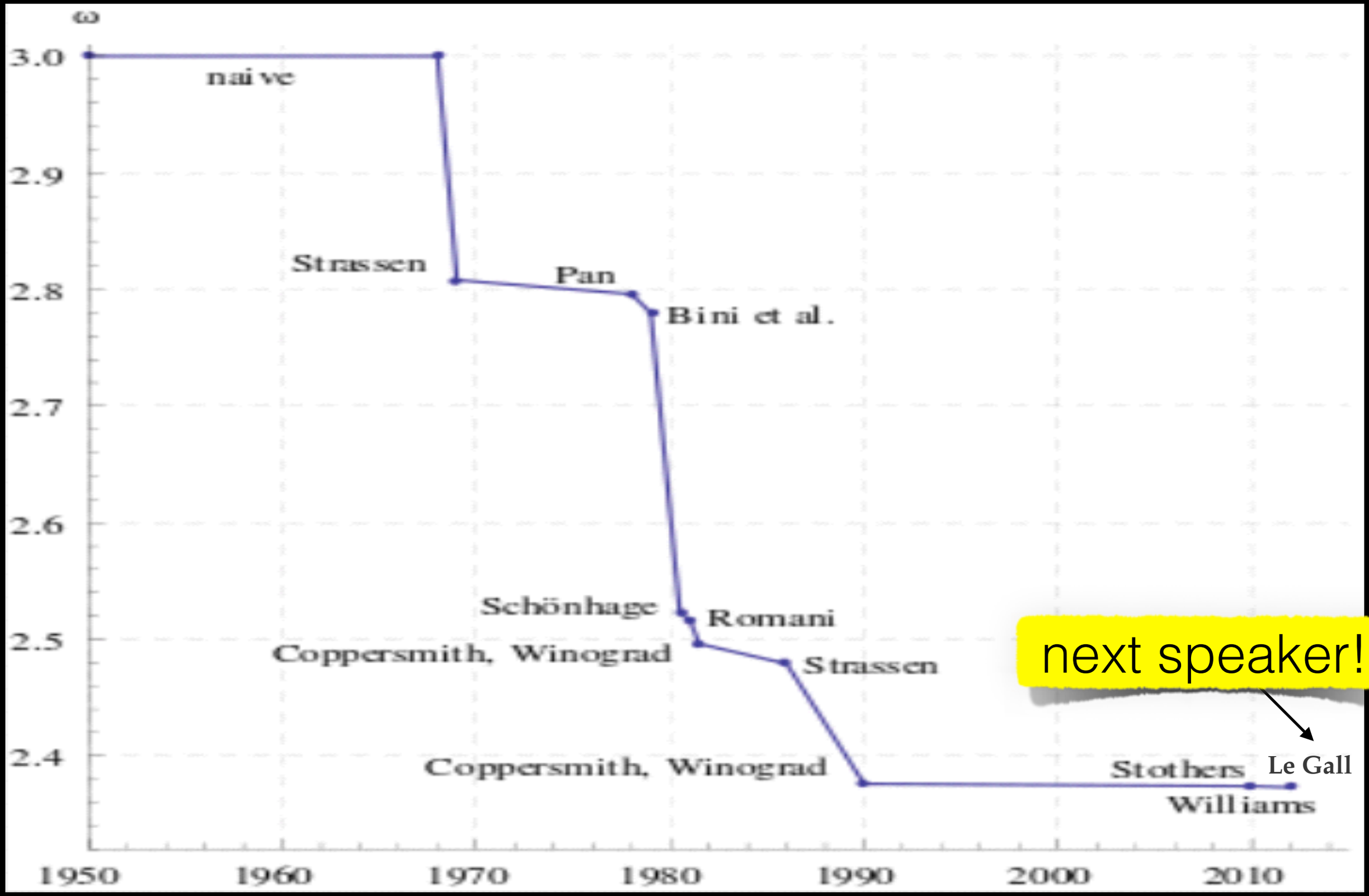
- **Problem**: Given a simple, undirected graph, what is the number of triangles T_3?
- Best known algorithm for sparse graphs runs in time $\sim m^{2\omega/(\omega+1)} = O(m^{1.41})$ where $\omega \leq 2.3727$ is the matrix multiplication exponent.

Triangle counting

m edges, n nodes



- **Problem**: Given a simple, undirected graph, what is the number of triangles T_3?
- Best known algorithm for sparse graphs runs in time $\sim m^{2\omega/(\omega+1)} = O(m^{1.41})$ where $\omega \leq 2.3727$ is the matrix multiplication exponent.
- In practice, simple triangle listing algorithms with running time $O(m^{1.5})$ are fastest.


ω
3.0
2.9
2.8
2.7
2.6
2.5
2.4
naive
Strassen
Pan
Bini et al.
Schönhage
Romani
Coppersmith, Winograd
Strassen
Coppersmith, Winograd
Stothers
Le Gall
Williams
next speaker!
1950
1960
1970
1980
1990
2000
2010

Graph streams

- We consider simple graphs that are too large to be loaded in memory.
- Streaming model: Only 1 pass over data allowed.
- Two input models:
  - incidence list streams: edges incident to each vertex arrive in succession (each one twice).
  - adjacency streams: edges arrive in any order.

Problem and known results

- **Goal**: Given stream of edge insertions/deletions, give an O(1)-approximation of T_3 with prob. 2/3.

Problem and known results

- **Goal**: Given stream of edge insertions/deletions, give an O(1)-approximation of T_3 with prob. 2/3.
- Manjunath et al., ESA ’11: $m^3/(T_3)^2$ space.

Problem and known results

- **Goal**: Given stream of edge insertions/deletions, give an O(1)-approximation of T_3 with prob. 2/3.
- Manjunath et al., ESA '11: $m^3/(T_3)^2$ space.
  - **New**: Optimal in terms of these parameters! http://arxiv.org/pdf/1404.4696v3.pdf

Problem and known results

- **Goal**: Given stream of edge insertions/deletions, give an O(1)-approximation of T_3 with prob. 2/3.
- Manjunath et al., ESA ’11: $m^3/(T_3)^2$ space.
  - **New**: Optimal in terms of these parameters! http://arxiv.org/pdf/1404.4696v3.pdf
- Ahn et al., SODA ’12: mn/T_3 space.

Problem and known results

- **Goal**: Given stream of edge insertions/deletions, give an O(1)-approximation of T_3 with prob. 2/3.

- Manjunath et al., ESA ’11: $m^3/(T_3)^2$ space.
  - **New**: Optimal in terms of these parameters!
    http://arxiv.org/pdf/1404.4696v3.pdf

- Ahn et al., SODA ’12: mn/T_3 space.

Both results: Very high update time.

2-path sampling algorithm

[Buriol et al. ’06]

- Assume the *incidence list* model, insertions only.
- **Idea**: Sample a 2-path and check whether it will be completed to a triangle later in the stream.
- *Transitivity coefficient* of graph G: $\alpha(G) = 3T_3/P_2$.

T_3 triangles, P_2 2-paths

2-path sampling algorithm

[Buriol et al. ’06]

- Assume the *incidence list* model, insertions only.
- **Idea**: Sample a 2-path and check whether it will be completed to a triangle later in the stream.
- *Transitivity coefficient* of graph G: $\alpha(G) = 3T_3/P_2$.

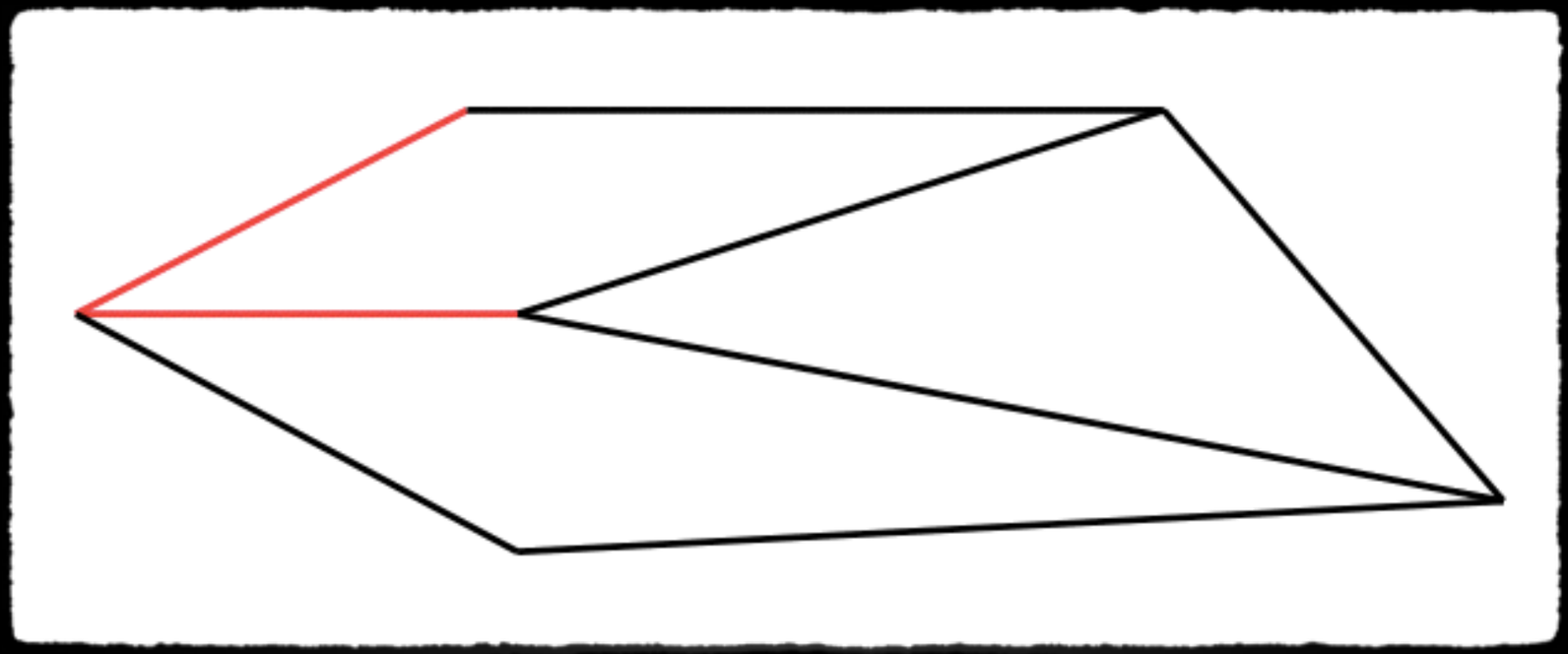
T_3 triangles, P_2 2-paths

- By sampling $O(1/\alpha(G))$ times we estimate $\alpha(G)$; incidence list streams: can compute P_2 exactly.

$T_3 = P_2\ \alpha(G)/3$

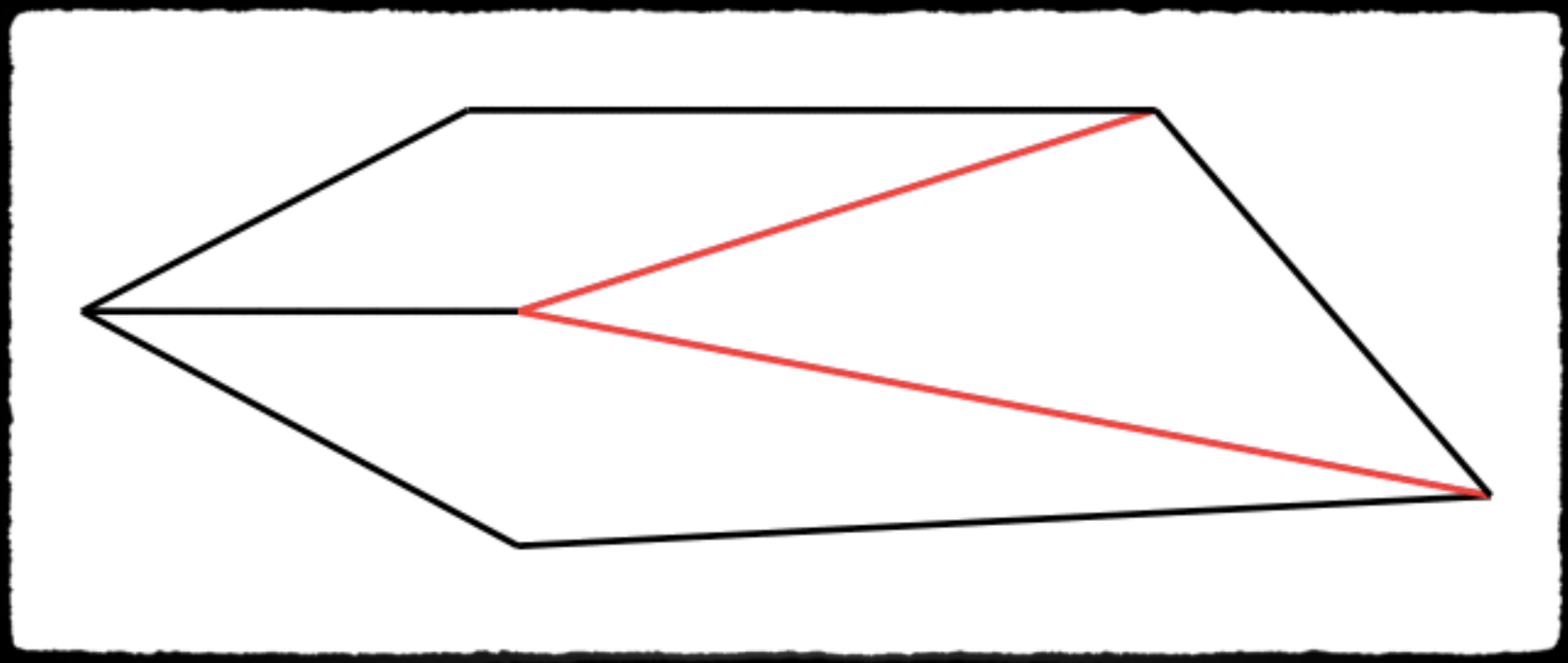
2-path sampling example

First sampled 2-path is not part of a triangle



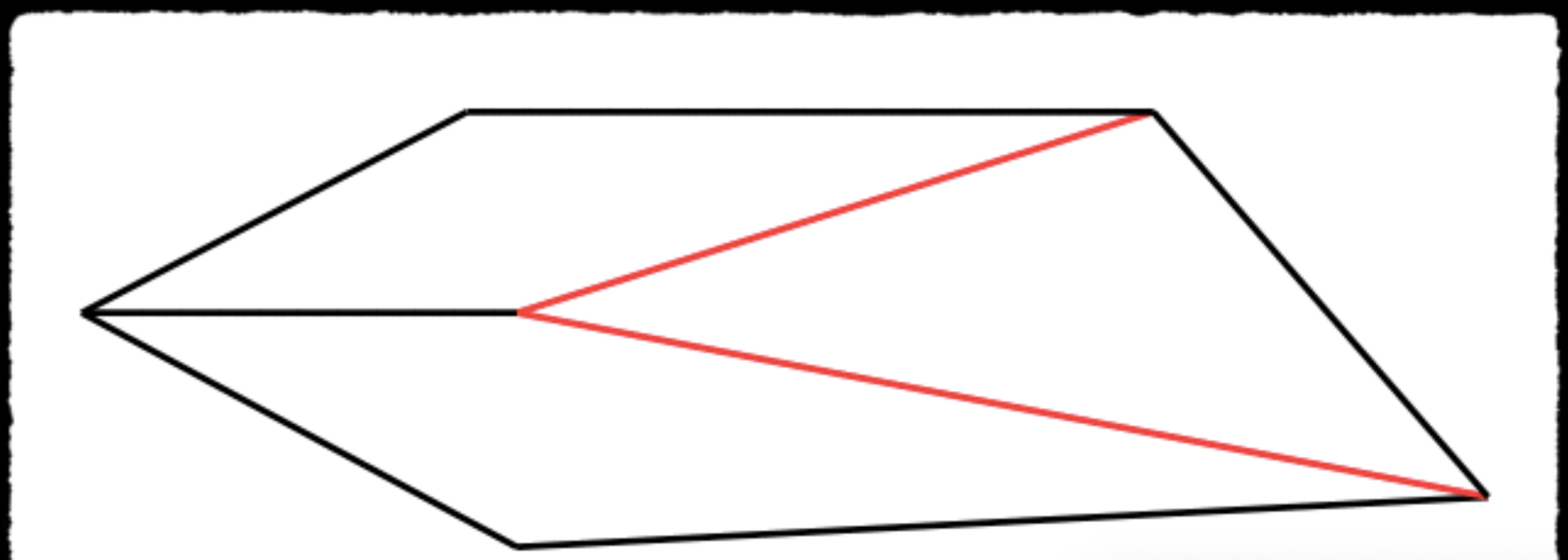
2-path sampling example

Second sampled 2-path **is** part of a triangle



2-path sampling example

Second sampled 2-path **is** part of a triangle



$P_2 = 14$

$\alpha(G) \approx 1/2$

$T_3 = P_2\, \alpha(G)/3 \approx 7/3$

Why edge deletion?

- Many real-life applications allow the deletion of edges.

- Most known algorithms for graph stream mining assume *insert-only* streams.
- **Here***:* General model where edges arrive in arbitrary order and can be deleted.
  - Problem: Cannot use this kind of sampling.

Edge sampling

[Tsourakakis et al. ’09, P.-Tsourakakis ’12]

- **Doulion algorithm**: Sample each edge with probability p; multiply number of triangles by p^{-3}.

Edge sampling

[Tsourakakis et al. ’09, P.-Tsourakakis ’12]

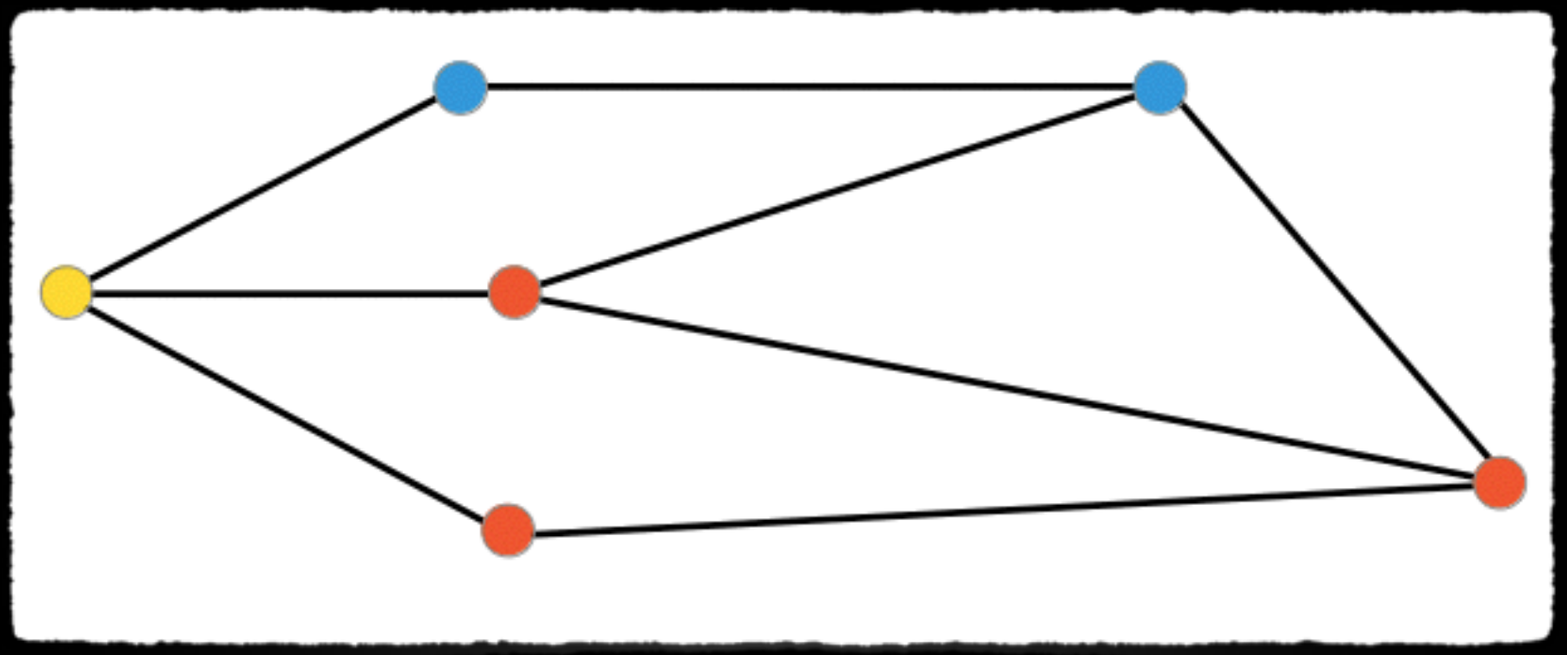
- **Doulion algorithm**: Sample each edge with probability p; multiply number of triangles by p^{-3}.
- **Colorful sampling**: Randomly color vertices with one of 1/p colors, sample edges whose endpoints have the same color; multiply number of triangles by p^{-2}.

Edge sampling

[Tsourakakis et al. ’09, P.-Tsourakakis ’12]

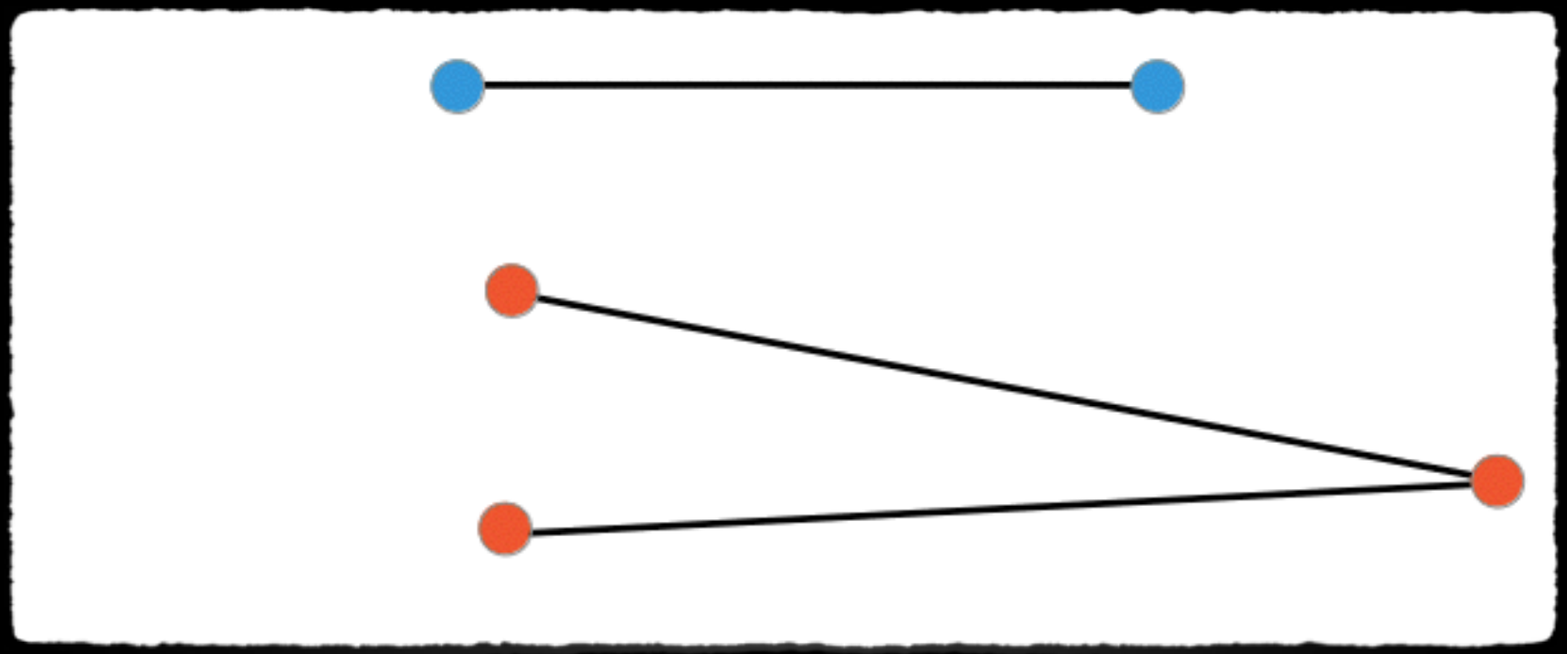
- **Doulion algorithm**: Sample each edge with probability p; multiply number of triangles by p^{-3}.
- **Colorful sampling**: Randomly color vertices with one of 1/p colors, sample edges whose endpoints have the same color; multiply number of triangles by p^{-2}.
  - Advantage: if we sample a 2-path, then we have also sampled any triangle it is part of.

Colorful sampling example



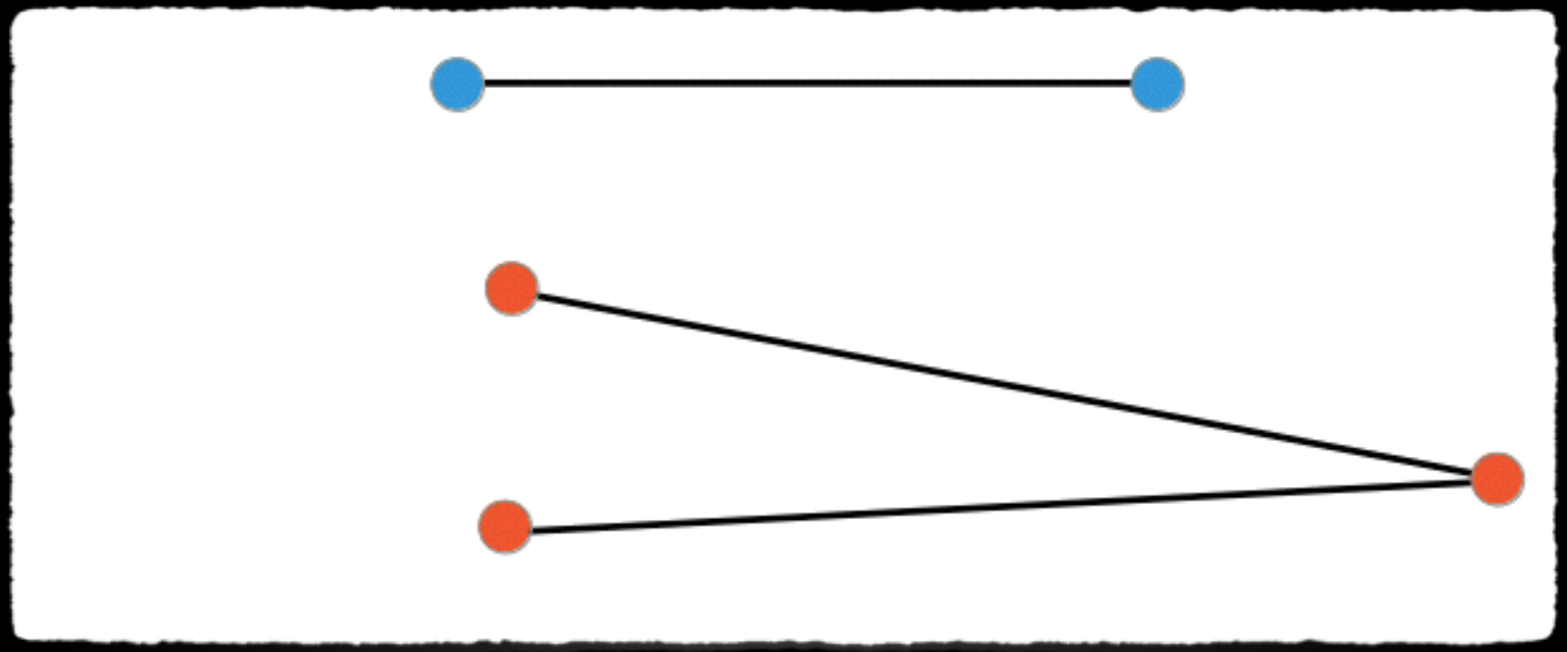
Colorful sampling example

No triangle sampled



Colorful sampling example

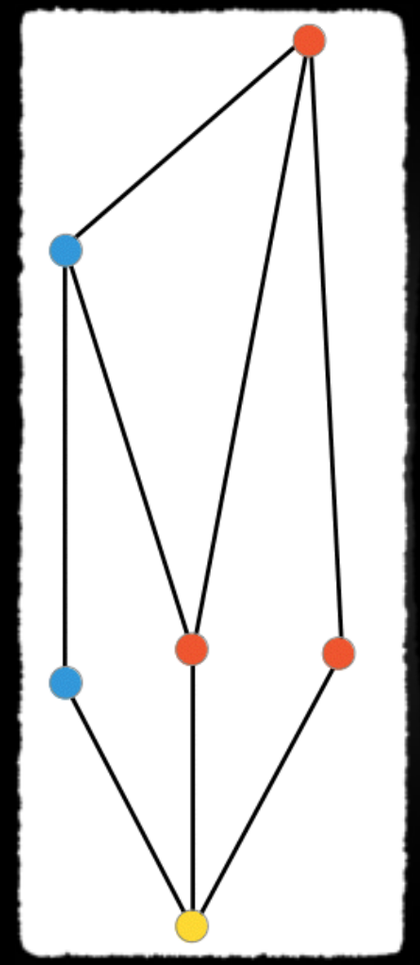
No triangle sampled



Estimate $T_3 \approx 0$

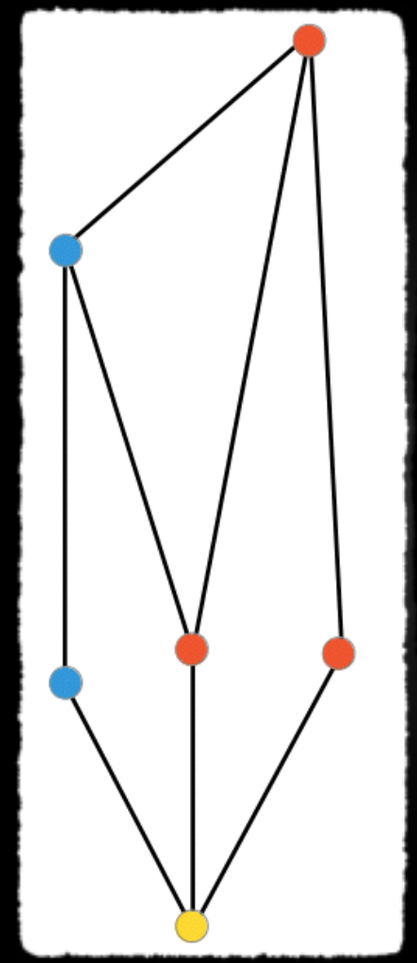
Combining the approaches

- Sample edges by colorful sampling.
- Choose random 2-path in the sample and check whether it is part of a triangle.
- By running several copies in parallel, estimate transitivity coefficient of G.
- In parallel, run a 2nd moment estimator to estimate the number of 2-paths in G.



Combining the approaches

- Sample edges by colorful sampling.
- Choose random 2-path in the sample and check whether it is part of a triangle.
- By running several copies in parallel, estimate transitivity coefficient of G.
- In parallel, run a 2nd moment estimator to estimate the number of 2-paths in G.



Central technical contribution:
Show that correlations among sampled 2-paths do not matter (too much) so we do get an estimate of α(G).

Main result

	Space
Ahn et al. [1]	$O(\frac{mn}{\epsilon^2 T_3} \log \frac{1}{\delta})$
Manjunath et al. [18]	$O(\frac{m^3}{\epsilon^2 T_3^2} \log \frac{1}{\delta})$
This work	$O(\frac{m}{\sqrt{b}\epsilon^3 \alpha} \log \frac{1}{\delta})$

$b = \max(n, P_2/n)$

Main result

	Update time	Space
Ahn et al. [1]	$O(n \log n)$	$O(\frac{mn}{\varepsilon^2 T_3} \log \frac{1}{\delta})$
Manjunath et	$O(\frac{m^3}{\varepsilon^2 T_3^2} \log \frac{1}{\delta})$	$O(\frac{m^3}{\varepsilon^2 T_3^2} \log \frac{1}{\delta})$
This work	$O(\frac{1}{\varepsilon^2 \alpha} \log \frac{1}{\delta})$	$O(\frac{m}{\sqrt{b}\varepsilon^3 \alpha} \log \frac{1}{\delta})$

$b = \max(n, P_2/n)$

Empirical study of graphs

Dataset	n	m	P_2/n	T_3	α	
Enron	36K	367K	710	727K	0.0853	
AS20000102	6.4K	13.2K	320	6.5K	0.0095	
Astro-Ph	18.7K	396.1K	681	1.35M	0.318	
Cond-Mat	23.1K	186.9K	85.5	173.3K	0.2643	
roadNet-CA	1.96M	5.53M	3.1	120.6K	0.0603	
DBLP	317K	1.05M	68	2.2M	0.3064	
Oregon	10.6K	22K	520	17.1K	0.0093	
Facebook	4K	88.2K	4.5K	1.61M	0.2647	
LiveJournal	4M	34.6M	1.5K	177.8M	0.1154	
Youtube	1.1M	2.9M	1.3K	3M	0.0062	
Amazon	334K	928K	30	667K	0.2052	
Skitter	1.7M	11.1M	9.5K	28.7M	0.0053	

Empirical study of graphs

n	P_2/n	T_3	α	m^3/T_3^2	mn/T_3	$m/(\sqrt{b}\alpha)$
7K	710	727K	0.0853	93.5K	**18K**	22.7K
2K	320	6.5K	0.0095	52K	**12.9K**	17.3K
6.1K	681	1.35M	0.318	34K	**5.5K**	9.1K
6.9K	85.5	173.3K	0.2643	215K	24.9K	**5K**
3M	3.1	120.6K	0.0603	11.55M	89.8M	**65.5K**
5M	68	2.2M	0.3064	239K	148K	**6.2K**
K	520	17.1K	0.0093	37.1K	**13.6K**	23.2K
2K	4.5K	1.61M	0.2647	264	**218**	4.97K
6M	1.5K	177.8M	0.1154	1.3M	778K	**147K**
M	1.3K	3M	0.0062	2.7M	1.06M	**467K**
8K	30	667K	0.2052	1.79M	464K	**5.6K**
1M	9.5K	28.7M	0.0053	1.66M	**657K**	1.6M

Open problems

- Our analysis requires a truly random coloring function. Give *explicit* hash function that works.
- **Conjecture**: In every graph with m edges and no isolated edge it is possible to find b=Ω(m) 2-paths that overlap (pairwise) in at most 1 vertex.
  - In the paper we show b > max(Ω(n), P_2/n). Showing conjecture will improve our space.

Thank you!