Graph sparsifiers

Charalampos (Babis) E. Tsourakakis\textsuperscript{1}

\textsuperscript{1}Boston University

CS591 Graph Analytics
What is a graph sparsifier?

- **Definition.** A sparsifier of a graph $G(V, E, w)$ is a sparse graph $H$ that is similar to $G$ in some useful notion.

- In the following:
  1. Triangle counts: **Triangle sparsifier**
  2. Densest subgraph: **Densest subgraph sparsifier**
  3. Shortest-path distances: **Spanner**
  4. Cut structure: **Cut sparsifier**
  5. Spectral properties: **Spectral sparsifier**
Triangle sparsifiers
Triadic closure

Source: [Easley-Kleinberg (fig.3.1, page 48)]
Triadic closure

If two people (B, C) have a common friend (A), then there is an increased likelihood that they will become friends in the future.

Why is it so common in social networks?

1. More likely B, C will meet
2. Trust/common interests
3. Reduction of social stress

BEARMAN, MOODY. 
Girls (teen) with low clustering coefficient may experience increased social stress.

HOMOPHILY.
SOCIOLOGICAL EXPLANATIONS
Triadic closure

How much more likely is that $B$ and $C$ will become friends if they have a friend in common?

- In the MSN Messenger that if two people have a common contact it is 18,000 times more likely that they are connected.

\[
\text{transitivity}(G) = \frac{3 \times \# \text{ triangles}}{\text{number of connected triples}}
\]

- Recall $\text{transitivity}(G) \in [0, 1]$ (why?) because it's a probability.
Exact triangle counting \( \text{(is not the same as listing).} \)

\[ G(V, E) \]

\[ \binom{n}{3} \]

- Clearly a naive cubic algorithm (i.e., \( O\left(\binom{n}{3}\right) \)) can list (and hence) count triangles.
- **Can we do better than naive?** Yes!
- Node-iterator
- Edge-iterator
  - Actually, we will show an improved triangle listing algorithm due to Chiba and Nishizeki, 1985 that runs in \( O(m\alpha(G)) \) time where \( \alpha(G) \) is the arboricity of the graph.
- Matrix multiplication
\[
A = \begin{pmatrix}
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 0
\end{pmatrix} \quad X = A^2 = \begin{pmatrix}
2 & 1 \\
1 & 1
\end{pmatrix}.
\]

\( (0, 1, 1) \cdot (1, 0, 1) = 1. \)
Exact triangle counting – Node iterator

Algorithm 3.2: TRIANGLE LISTER "node-iterator"

Input: graph \( G = (V, E) \), arbitrary order \(<\) of the nodes
Output: all triangles of \( G \)

\[
\text{for } v \in V \text{ do}
\]

\[
\text{for all pairs of Neighbors } \{u, w\} \text{ of } v \text{ do}
\]

\[
\text{if } \{u, w\} \in G \text{ then}
\]

\[
\text{if } u < v < w \text{ then}
\]

\[
\text{output triangle } \{u, v, w\};
\]

Source: [Algorithmic aspects of triangle-based network analysis, T. Schank]

\[
\sum_{v \in V} \left( \frac{\deg(v)}{2} \right) \leq n \cdot \max_{v \in V} \left( \frac{\deg(v)}{2} \right)
\]

\[
= O(n \cdot \Delta(G)).
\]

Runtime?
Exact triangle counting – Matrix Multiplication

- **We will describe a simple algorithm based on matrix multiplication.**
- **There exists a more sophisticated algorithm due to Alon, Yuster, Zwick** *Finding and counting given length cycles*
- **Let** $A$ **be the adjacency matrix of graph** $G$.
  1. Compute $X = A^2$
  2. Check for each pair $(i, j)$ if $a_{ij}x_{ij} > 0$. If yes, a triangle was found!
- **Why does it work?**
- **What is the running time?**
Exact triangle counting – Matrix Multiplication

**Exercise:** $A^3$\[ t = \text{Tr}(A^3). \]

- Let $A$ be the adjacency matrix of graph $G$.

1. Compute $X = A^2$
2. Check for each pair $(i, j)$ if $a_{ij}x_{ij} > 0$. If yes, a triangle was found!

$$X = \sum_{k=1}^{n} a_{ik} \cdot a_{kj}$$

$X$ is NOT PRACTICAL.

- Simple undirected graph $G = \{1, \ldots, n\}$.

Triangles edge $(i, j)$ is contained in $J$ common neighbors between $i$ and $j$.

**Strassen’s Algorithm:** $O(n \log_27)$

$O(n^3) = O(n^{\omega})$

$\omega = 2.371$ (due to Virginia V. MIT).

D & C subcubic!
**Exact triangle counting – Arboricity**

**Dfn.** Arboricity of graph $G$ is defined to be the minimum number of edge-disjoint spanning forests into which $G$ can be decomposed.

**Example decomposition $o(G)$**

$$w_c = 5$$

$$\varepsilon((1, 6, 2)) = 2$$

Here, happens to have spanning trees.
Remark: It is trivial to see $\alpha(G) \leq m$.

\[ m \geq n-1. \]

\[
\alpha(G) = \max_{H \subseteq G} \left\lfloor \frac{e(H)}{|H| - 1} \right\rfloor
\]

Nash-Williams, 60s

Exercise.

\[ \alpha(G) \leq \frac{2m + n}{2} \]

\[ \alpha(H) = O(m) \]
procedure K3(G);
{Let $G$ be a graph with $n$ vertices and $m$ edges.}
begin
sort the vertices $v_1, v_2, \ldots, v_n$ of $G$ in such a way that $d(v_1) \geq d(v_2) \geq \ldots \geq d(v_n);$  
for $i = 1$ to $n - 2$
do begin 
{find all the triangles containing vertex $v_i$, each of which corresponds to an edge joining two neighbours of $v_i$.}
1: mark all the vertices adjacent to $v_i;$
for each marked vertex $u$
do begin 
2: for each vertex $w$ adjacent to $u$
do if $w$ is marked 
thend print out triangle $(v_i, u, w);$ 
3: erase the mark from $u$
end;
{delete $v_i$ from $G$ so that no duplication occurs.}
4: delete vertex $v_i$ from $G$ and let $G$ be the resulting graph
end
end;
Algorithm K3 lists all triangles in $G$ in time $O(m\alpha(G))$.

- It is clear that the running time is
  $$O(n + m) + O\left(\sum_{e=(u,v)} \min(\deg(u), \deg(v))\right).$$

- We get the desired bound by using the following lemma (see HW2):
  $$\sum_{e=(u,v)} \min(\deg(u), \deg(v)) \leq 2\alpha(G)m.$$
Triangle Sparsifiers

- Triangle sparsifiers were originally proposed in the paper “Doulion: counting triangles in a massive graph with a coin” by Tsourakakis et al.
- A tighter analysis followed in the paper “Triangle sparsifiers” by Tsourakakis, Kolountzakis, Miller.
\[
\frac{1}{N(e,f)} \sum Y_{e,f,g} = p^3 t \alpha
\]

**Key idea.**

\[
\Pr(\Delta(e, f, g) \text{ survives}) = \begin{cases} 
1 & \text{if } N(e, f, g) \text{ survived} \\
0 & \text{ otherwise}
\end{cases}
\]

\[E(t_{4\ell}) = t_{4\ell} \cdot p^3 \cdot \alpha\]

\[
E(\# \text{triangle in } H) = p^3 t_{4\ell} = p^3 t_{\alpha}
\]

**Algorithm.**

- Toss a coin for every edge. With probability \( p \)
- Count \( \Delta \) as in \( H \).
- Return \( t_{4\ell} / p^3 \)

**Key question.** How small can \( p \) be so that \( \alpha \geq t_{\alpha} \)?

**Concentration.** But Chernoff won't work (indicator RV)

\[\text{edges } e, f, g \text{ ALL survive.}\]
Triangle Sparsifiers

- Triangle sparsifiers were originally proposed in the paper “Doulion: counting triangles in a massive graph with a coin” by Tsourakakis et al.
- A tighter analysis followed in the paper “Triangle sparsifiers” by Tsourakakis, Kolountzakis, Miller.

**Require:** Set of edges $E \subseteq \binom{[n]}{2}$

{Unweighted graph $G([n], E)$}

**Require:** Sparsification parameter $p$

Pick a random subset $E'$ of edges such that the events \{e $\in E'$\}, for all e $\in E$ are independent and the probability of each is equal to $p$.

$t' \leftarrow$ count triangles on the graph $G'([n], E')$

Return $T \leftarrow \frac{t'}{p^3}$
Do we need to toss a coin for each edge?

**Claim.** The edge sampling can run in $O(pm)$ expected time.

- We do not wish to “toss a $p$-coin” $m$ times in order to construct $E'$.
- Sublinear sampling: if $p$ is small, then we toss significantly less than $m$ coins!
- **How can we achieve this?**
- **Observation.** The number of unsuccessful events, i.e., edges which are not selected in our sample, until a successful one follows a **geometric distribution**.
- **Key idea:** Generate these numbers assuming access to uniform RV in $[0,1]$ (Homework 2, problem).

```python
import numpy as np
np.random.uniform()
```
\[ m = 10^9 \]

\[ e_1, e_2, e_3, e_4, e_5, \ldots, e_{m-1}, e_m \]

\[ 0, 0, 0, 0, 0, \ldots, 1 \]

\[ \color{orange} \text{NAIVELY} \]

\[ 10^9 \text{ tosses} \]

\[ \hat{p} = \frac{\log(m)}{m} = \frac{9}{10^9} \]

\( p \) is small.

\[ 10^9 \text{ tosses to get in expectation} \]

9 times heads!

Really wasteful!

\( p = 0 \) I would not have to toss any coin!

WHAT if \( \frac{1}{100} p \) is small?

geometric R.V. \( (p) (1) \)

1)

1)

1)

1)

1)

1)

1)

1)

\( a) \) if access to \( \text{UNIFORM} \)
Triangle Sparsifiers

- Let $X_e = 1$ or $0$ depending on whether the edge $e$ of graph $G$ survives in $G'$.

- Then

$$T = \sum \Delta(e, f, g) X_e X_f X_g$$

where $\Delta(e, f, g) = 1$ (edges $e$, $f$, $g$ form a triangle).

- What is the expectation $E[T]$?

$$E(T) = \frac{\rho^3}{p^3} \rightarrow \# \Delta s \text{ in } G.$$

- **How small can $\rho$ be, and still guarantee concentration?**

\[ \text{L} \rightarrow \text{How much can we sparsify} \]
\[
T = X_{12} X_{13} X_{23} + X_{45} X_{46} X_{56}.
\]
Kim-Vu inequality

- Kim-Vu inequality allows under certain conditions to derive strong concentration when the polynomial of interest is not smooth.

\[ T = X_{12}X_{13}X_{23} + X_{12}X_{14}X_{24} + \cdots + X_{12} \cdots X_{1,n-2}X_{2,n-2}. \]

If \( X_2 \) flips (0 \( \rightarrow \) 1 or 1 \( \rightarrow \) 0), then \( T \) changes by \( n-2 = O(n) \).
Kim-Vu inequality 

- Kim-Vu inequality allows under certain conditions to derive strong concentration when the polynomial of interest is not smooth.
- Let $Y = Y(t_1, \ldots, t_m)$ be a positive polynomial of $m$ Boolean variables $[t_i]_{i=1..m}$ which are independent.
- $Y$ is totally positive if all of its coefficients are non-negative variables.
- $Y$ is homogeneous if all of its monomials have the same degree and we call this value the degree of the polynomial.

\[
Y = 16t_1t_2 + 108t_1t_2t_3
\]
Kim-Vu inequality

- $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{Z}_+^m$
- Define the partial derivative
  \[ \partial^\alpha Y = \left( \frac{\partial}{\partial t_1} \right)^{\alpha_1} \cdots \left( \frac{\partial}{\partial t_m} \right)^{\alpha_m} Y(t_1, \ldots, t_m) \]
- Denote by $|\alpha| = \alpha_1 + \cdots + \alpha_m$ the order of $\alpha$.
- For any order $d \geq 0$, define
  \[ \mathbb{E}_d(Y) = \max_{\alpha: |\alpha| = d} \mathbb{E}(\partial^\alpha Y) \]
  and
  \[ \mathbb{E}_{\geq d}(Y) = \max_{d' \geq d} \mathbb{E}_{d'}(Y). \]
Kim-Vu inequality

**Theorem (Kim-Vu)**

There is a constant $c_k$ depending on $k$ such that the following holds. Let $Y(t_1, \ldots, t_m)$ be a totally positive polynomial of degree $k$, where $t_i$ can have arbitrary distribution on the interval $[0, 1]$. Assume that:

\[ \mathbb{E}[Y] \geq \mathbb{E}_{\geq 1}(Y) \quad (2) \]

Then for any $\lambda \geq 1$:

\[ \Pr \left[ |Y - \mathbb{E}[Y]| \geq c_k \lambda^k (\mathbb{E}[Y] \cdot \mathbb{E}_{\geq 1}(Y))^{1/2} \right] \leq e^{-\lambda + (k-1) \log m} \quad (3) \]
Theorem (Tsourakakis, Kolountzakis, Miller)

Suppose $G$ is an undirected graph with $n$ vertices, $m$ edges and $t$ triangles. Let also $\Delta$ denote the size of the largest collection of triangles with a common edge. Let $G'$ be the random graph that arises from $G$ if we keep every edge with probability $p$ and write $T$ for the number of triangles of $G'$. Suppose that $\gamma > 0$ is a constant and

\[
\frac{pt}{\Delta} \geq \log^{6+\gamma} n, \text{ if } p^2 \Delta \geq 1,
\]

and

\[
p^3 t \geq \log^{6+\gamma} n, \text{ if } p^2 \Delta < 1.
\]

for $n \geq n_0$ sufficiently large.
Theorem ((cont.) Tsourakakis, Kolountzakis, Miller)

Then

\[ \Pr \left[ |T - \mathbb{E}[T]| \geq \epsilon \mathbb{E}[T] \right] \leq n^{-K} \]

for any constants \( K, \epsilon > 0 \) and all large enough \( n \) (depending on \( K, \epsilon \) and \( n_0 \)).
Triangle sparsifiers in practice

• Before we prove the theorem, observe that the optimal value of $p$ depends on $t$.

• Let’s see how one may apply it in practice.

• The next slide shows an example for the Wikipedia 2005/09.
  • $n = 1\,634\,989$, $m = 18\,540\,603$, $t = 45\,542\,697$.

• Doubling trick – concentration for $p = 0.02$.

• Speedups reported are over node iterator and averaged over the ten experiments.
# Triangle sparsifiers in practice – Doubling trick

<table>
<thead>
<tr>
<th>$p$</th>
<th>Ratios</th>
<th>Sparsification (secs)</th>
<th>Average Speedup (x$\times$ faster)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>$0.9442, 1.4499$</td>
<td>8</td>
<td>$7090$</td>
</tr>
<tr>
<td></td>
<td>$1.14, 1.37$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>$0.9112, 1.0183$</td>
<td>8.64</td>
<td>$2880$</td>
</tr>
<tr>
<td></td>
<td>$0.8975, 0.9579$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.9716, 0.9771$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.9524, 0.9551$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.9606, 0.9716$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>$1.0043, 1.0336, 1.0035$</td>
<td>8.65</td>
<td>$1500$</td>
</tr>
<tr>
<td></td>
<td>$0.9791, 1.0222$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.9865, 0.9816$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td>$0.9895, 1.018$</td>
<td>8.58</td>
<td>$825$</td>
</tr>
<tr>
<td>0.05</td>
<td>$0.9979, 0.9716$</td>
<td>9.84</td>
<td>$402$</td>
</tr>
</tbody>
</table>
Triangle sparsifiers – Proof

\[ t \left[ \frac{\partial T}{\partial X_e} \right] = \sum_{\Delta(a,f,g)} E(X_e X_g) = p^2 \Delta(e) \]

\[ E_4(T) = p^2 \Delta \left( \Delta = \max_e \Delta(e) \right) \]

\[ E \left[ \frac{\partial^2 T}{\partial X_e \partial X_f} \right] = p \cdot \mathbb{I}(f \in \Delta(e, f, g)) \]

Hence, \( E_2(T) \leq p \). Obviously \( E_3(T) \leq 1 \).
Triangle sparsifiers – Proof

\[ E_{\geq 3} (T) \leq 1 \]

\[ E_{\geq 2} (T) \leq \left( \frac{1}{\phi} \right)^{\frac{1}{\phi}} \] \quad (\phi \leq 1)

\[ E_{\geq 1} (T) \leq \max \left\{ 1, \phi^2 \Delta^2 \right\} \quad \left( E_{\leq 0} (T) = \phi^3 t \right) \]

\[ E_{\geq 0} (T) = \max \left\{ 1, \phi^2 \Delta, \phi^3 t \right\} \]

**CASE 1**

if \[ \phi^2 \Delta < 1 \]

\[ E_{\geq 1} (T) \leq 1 \]

and therefore.

\[ E_{\geq 0} (T) = \phi^3 t \]

By condition.

\[ \phi^3 t \geq \log G_\Delta(n) \] if \( \phi^2 \Delta < 1 \)
Triangle sparsifiers – Proof

\[ \text{CASE 2} \quad (p^2 \Delta \geq 1) \]

\[ E_{\geq 1}(T) \leq p^2 \Delta \quad \text{AND from condition} \]

\[ p^t \cdot \log \frac{648(n)}{\Delta} \quad \text{we get} \quad E_{\geq 20}(T) = p^3 t. \]

By \( K14^{-TVU} \), we get for some constant \( c_3 > 0 \)

\[ \Pr \left[ \left| T - E(T) \right| > c_3 \lambda^3 \left( E(T) E_{\geq 1}(T) \right)^{1/2} \right] \leq e^{-2t + 2 \log n} \]
Notice in both cases, \( \mathcal{F}(T) \geq \mathcal{E}_{\Delta_1}(T) \).

In case 1, we choose 
\[
\lambda = \frac{\epsilon^{\frac{n}{3}}}{C_3^{\frac{n}{3}}} \left( \frac{p^3 t}{\Delta} \right)^{\frac{n}{6}}.
\]

In case 2, we set 
\[
\lambda = \frac{\epsilon^{\frac{n}{3}}}{C_3^{\frac{n}{3}}} \left( \frac{p^3 t}{\Delta} \right)^{\frac{n}{6}}.
\]

From our two assumptions, \( \lambda = (k+2) \log n \). (Verify).

Pr. \( |T - \mathcal{E}(T)| \geq \epsilon \mathcal{E}(T) \) \( \leq \epsilon^{-k} \).
IN PRACTICE, the optimal value of $p$ depends on $k!$

The quantity we want to estimate.

Doubling Trick.

Start with $p$ really low.

Perform 5-10 sparsifications.

If "concentration"

Stop and output $\text{avg.}$

otherwise $p \leftarrow 2p$
Triangle sparsifiers – Proof

**Corollary:** if \( t = \omega(n^{3/2}) \), \( \Delta = O(n) \), then \( p = \frac{1}{\sqrt{n}} \)

Results in concentration.

Suppose we use `NODE_ITERATOR`, HENCE the black-box complexity is

\[
\sum_{v \in V} \left( \frac{\deg(v)}{\sqrt{n}} \right).
\]

We can justify \( O(n) \) speedups! \( O\left(\frac{1}{p^2}\right) \).
Can we lower the degree of the multivariate polynomial?

**Require:** Unweighted graph $G([n], E)$

**Require:** Number of colors $N = 1/p$

Let $f : V \rightarrow [N]$ have uniformly random values $E' \leftarrow \{\{u, v\} \in E \mid f(u) = f(v)\}$

$T \leftarrow$ number of triangles in the graph $(V, E')$

return $T/p^2$

• However, we lose independence...

• Nonetheless we can prove the following theorem:
Theorem (Pagh-Tsourakakis)

Let $n$, $t$, $\Delta$, $T$ denote the number of vertices in $G$, the number of triangles in $G$, the maximum number of triangles an edge of $G$ is contained and the number of monochromatic triangles in the randomly colored graph respectively. Also let $N = \frac{1}{p}$ the number of colors used. If $p \geq \max \left( \frac{\Delta \log n}{t}, \frac{\log n}{\sqrt{t}} \right)$, then $T \sim \mathbb{E}[T]$ with probability $1-o(1)$.
From Chebyshev's inequality, if \( \text{var}(T) = o(E(T)^2) \), I have concentration. (\( T \sim E(T) \) w.p. \( 1 - o(1) \)).

For the \( i \)-th triangle, I will define \( X_i = 1 \) if it's monochromatic. (\( X_1, X_2, \ldots, X_t \)).

\[
T = \sum_{i=1}^{t} X_i
\]

\[
P(T_i = 1) = \left( \frac{1}{N^2} \right) = p^2.
\]
Colorful triangle counting – Proof

\[ E(T) = p^2 \cdot t. \]

\[ C = \sum_{i,j} \text{Pr.} (X_i \land X_j). \]

\[ \text{VAR}(T) \leq E(T) + C. \]

\[ \leq p^2 t + p^3 \sum_{E \in E} \Delta^2(E) \leq p^2 t + 3p^3 \cdot k \Delta. \]
How can we guarantee concentration?

\[ \var(T) = \text{Var}(T) = o\left( \frac{E(T)^2}{\text{Var}(X_i) \cdot E(X_i)} \right) \]

\[ T = \left( \sum X_i \right)^2 / \text{Var}(X_i) \cdot E(X_i) \]

\[ \text{Var}(T) = \frac{E(T^2) - (E(T))^2}{E(T)} \]

\[ E(T^2) = \sum E(X_i^2) + \sum E(X_i X_j) \]

\[ \text{Var}(T) = \sum \text{Var}(X_i) + \sum \text{Cov}(X_i, X_j) \]

\[ \left( \frac{E(T^2)}{E(T)} \right)^2 = p^4 + t^2 \Rightarrow p^2 + t^2 + \left( 3p^3 + t \Delta \right) \approx \Rightarrow p^2 t \gg 1 + 3p \Delta \]
Colorful triangle counting – Proof

CASE 2

\[ p^t = \hat{\omega}(n), \quad \hat{\omega}(n) \text{ is any slowly growing func.} \]

\[ \log^2(n) \]

\[ p = \frac{\log n}{t} \]

CASE III

\[ p \Delta \geq \frac{1}{3} \]

Suffices to set \[ \frac{pt}{\Delta} = \log n \]
if \( p \geq \max \left( \frac{\Delta \log n}{t}, \frac{\log n}{\sqrt{t}} \right) \)

then \( \text{VAR}(T) = o((E(T))^2) \).

By Chebyshev, \( T \sim E(T) \) wp. 1 - o(1).