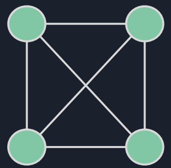
ProbGraph: High-Performance and High-Accuracy Graph Mining with Probabilistic Set Representations

Paper by Besta et al.
Presentation by Xander Morgan

Recurring theme: operating on sets of vertices

Example: count number of 4-cliques (same idea as counting triangles)



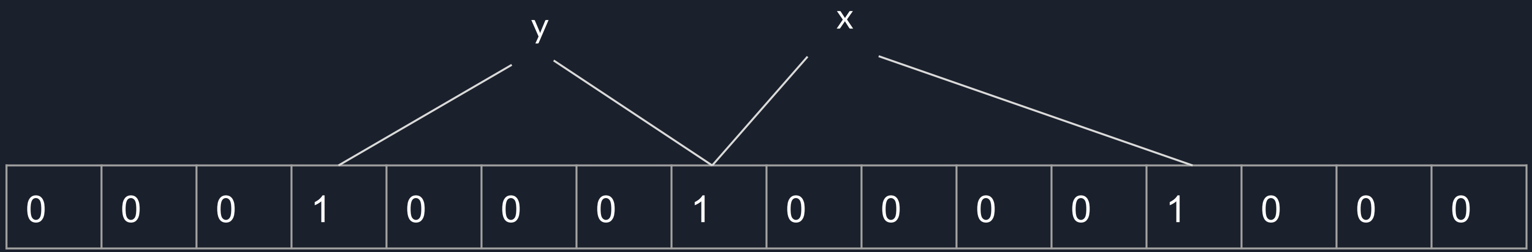
Main idea: Use probabilistic set representations to approximate!

```
/* Input: A graph $G$. Output: Number of 4-cliques $ck \in \mathbb{N}$. */
/Derive a vertex order $R$ s.t. if $R(v) < R(u)$ then $d_v \leq d_u$:
for $v \in V$ [in par] do: $N_v^+ = \{u \in N_v | R(v) < R(u)\}$
$ck$ = 0;
for $u \in V$ [in par] do:
  for $v \in N_u^+$ [in par] do:
    $C_3$ = $N_u^+ \cap N_v^+$ //Find 3-cliques
    for $w \in C_3$ do: //For each 3-clique...
      $ck$ += $|N_w^+ \cap C_3|$ //Find 4-cliques
```

Bloom Filter

- Key idea: use bit vector to approximate set
- Parameterized by length l and b hash functions $h_1, \ldots, h_b$

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



MinHash

k-Hash variant:

- Have k independent hash functions $h_1, \ldots, h_k$
- For each hash, retain an element with a minimum hash for a total of k retained elements
- Sampling WITH replacement
- Multiset

1-Hash variant

- Have a single hash function
- Retain k elements with k smallest hashes
- Sampling WITHOUT replacement

Estimation background

- Biased coin, heads with probability p, but we don't know what p is.
- Flip the coin n times. How to estimate p given data X_1, X_2, … X_n?

$$\hat{p} = \arg\max_p \prod_{j=1}^{n} p^{x_j} (1-p)^{1-x_j} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Estimation background

- Let $s = p^2$
- How to estimate s?

$$\hat{s} = \arg\max_s \prod_{j=1}^{n} \left(\sqrt{s}\right)^{x_j} \left(1 - \left(\sqrt{s}\right)\right)^{1-x_j} = \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2$$

Estimation background

- Let $s = p^2$
- How to estimate s?
- Bias in this estimator, but still MLE

$$\mathbb{E}[\hat{S}] = p^2 + \frac{1}{n}p(1-p) > p^2 \text{ for } p \in (0,1)$$

Bloom Filter Estimation

Length of bloom filter (in bits)

Number of hashes used

$$\widehat{|X|}_S = -\frac{B_X}{b} \log\left(1 - \frac{B_{X,1}}{B_X}\right)$$

Number of "1" bits in the bloom filter

$$\widehat{|X \cap Y|}_{AND} = -\frac{B_{X\cap Y}}{b} \log\left(1 - \frac{B_{X\cap Y,1}}{B_{X\cap Y}}\right)$$

Bloom Filter Estimation

Weakness: This is not a very useful proposition because we don't know the number of 1's in the bloom filter for the intersection. The authors note this, but they resolve with an unsatisfying/not-fully-motivated approximation (bitwise and).

Proposition IV.1. *Let* $\widehat{|X \cap Y|}_{AND}$ *be the estimator defined in Eq. (2). For* $B_{X\cap Y}, b \in \mathbb{N}$ *such that* $b = o(\sqrt{B_{X\cap Y}})$, *and a set* $X \cap Y$ *such that* $b|X \cap Y| \leq 0.499 B_{X\cap Y} \cdot \log B_{X\cap Y}$ *the following holds:*

$$E\left[\left(\widehat{|X \cap Y|}_{AND} - |X \cap Y|\right)^2\right] \leq$$

$$(1 + o(1))\left(e^{|X\cap Y|b/(B_{X\cap Y}-1)}\frac{B_{X\cap Y}}{b^2} - \frac{B_{X\cap Y}}{b^2} - \frac{|X \cap Y|}{b}\right)$$

k-Hash estimation

$$|M_X \cap M_Y| \sim Bin(k, J_{X,Y})$$

- Need to be careful about multiset intersections. (Discuss example)
- MLE estimate

$$J_{X,Y} = |X \cap Y| / |X \cup Y|$$

$$|X \cup Y| = |X| + |Y| - |X \cap Y|,$$

$$|\widehat{X \cap Y}|_{kH} = \frac{\widehat{J_{X,Y}}_{kH}}{1 + \widehat{J_{X,Y}}_{kH}} (|X| + |Y|)$$

$$\widehat{J_{X,Y}}_{kH} = \frac{|M_X \cap M_Y|}{k}$$

k-Hash estimation

Proposition IV.2. *Let* $|\widehat{X \cap Y}|_{kH}$ *be the estimator from Eq. (5). Then, an upper bound for the probability of deviation from the true* $|X \cap Y|$, *at a given distance* $t \geq 0$, *is:*

$$P\left(\left||\widehat{X \cap Y}|_{kH} - |X \cap Y|\right| \geq t\right) \leq 2e^{-\frac{2\ k\ t^2}{(|X|+|Y|)^2}} \tag{6}$$

1-Hash estimation

$$\widehat{J_{X,Y}}_{1H} = \frac{|M^1_X \cap M^1_Y|}{k}$$

$$|\widehat{X \cap Y}|_{1H} = \frac{\widehat{J_{X,Y}}_{1H}}{1+\widehat{J_{X,Y}}_{1H}}(|X| + |Y|)$$

1-Hash estimation

Proposition IV.3. *Consider* $|\widehat{X \cap Y}|_{1H}$. *Then, an upper bound for the probability of deviation from the true intersection set size, at a given distance* $t \geq 0$, *is:*

$$P\left(\left||\widehat{X \cap Y}|_{1H} - |X \cap Y|\right| \geq t\right) \leq 2e^{-\frac{2\ k\ t^2}{(|X|+|Y|)^2}}$$

Estimation summary

Result	Where	Class	AU	CN	ML	IN	AE
$\widehat{\|X\|}_S$	Eq. (1)	BF	👍★	👍★	✖	✖	✖
$\widehat{\|X \cap Y\|}_{AND}$ ★	Eq. (2)	BF	👍★	👍★	✖	✖	✖
$\widehat{\|X \cap Y\|}_{L}$ ★	§ IV-B	BF	👍★	👍★	✖	✖	✖
$\widehat{\|X \cap Y\|}_{kH}$	Eq. (5)	k-Hash	👍★	👍★	👍★	👍★	👍★
$\widehat{\|X \cap Y\|}_{1H}$	§ IV-D	1-Hash	👍★	👍★	✖	✖	✖

TABLE II: Summary of theoretical results (estimators) related to $\widehat{|X|}$ and $\widehat{|X \cap Y|}$. "★": a new result provided in this work (a new estimator or proving a certain novel property of a given estimator). "**CN**": a consistent estimator. "**AU**": an asymptotically unbiased estimator. "**ML**": an MLE estimator. "**IN**": an invariant estimator. "**AE**": an asymptotically efficient estimator.

Estimation summary

Result	Where	Class	Q	MS	CO
$\widehat{\lvert X \rvert}_S$ ★	Eq. (1)	BF	P ★	👍	👍
$\widehat{\lvert X \cap Y \rvert}_{AND}$ ★	Eq. (3)	BF	P ★	👍	👍
$\widehat{\lvert X \cap Y \rvert}_L$ ★	§ IV-B	BF	P ★	👍	👍
$\widehat{\lvert X \cap Y \rvert}_{kH}$ ★	Eq. (6)	k-Hash	E ★	✖	👍
$\widehat{\lvert X \cap Y \rvert}_{1H}$ ★	Eq. (7)	1-Hash	E ★	✖	👍

TABLE III: Summary of theoretical results (bounds) related to $\widehat{\lvert X \rvert}$ and $\widehat{\lvert X \cap Y \rvert}$. "★": a new result provided in this work. "**Q**": the quality of a given bound, "**P**": polynomial, "**E**": exponential. "**MS**": an MSE bound. "**CO**": a concentration bound.

Application of ideas

- Store graph in CSR format

- Parameter $0 \leq s \leq 1$ to choose how much extra storage to use for PG estimators

- Bloom filter is a bitvector (no surprise)

- Min-Hash are series of integers

Application of ideas

	CSR	PG (BF)	PG (MH)
Triangle Counting (work):	$O\left(nd^2\right)$	$O\left(\frac{ndB_X}{W}\right)$	$O\left(ndk\right)$
Triangle Counting (depth):	$O\left(\log d\right)$	$O\left(\log\left(\frac{B_X}{W}\right)\right)$	$O\left(\log k\right)$
4-Clique Counting (work):	$O\left(nd^3\right)$	$O\left(\frac{nd^2B_X}{W}\right)$	$O\left(nd^2k\right)$
4-Clique Counting (depth):	$O\left(\log^2 d\right)$	$O\left(\log d \log\left(\frac{B_X}{W}\right)\right)$	$O\left(\log^2 k\right)$
Clustering (work):	$O\left(nd^2\right)$	$O\left(\frac{ndB_X}{W}\right)$	$O\left(ndk\right)$
Clustering (depth):	$O\left(\log d\right)$	$O\left(\log\left(\frac{B_X}{W}\right)\right)$	$O\left(\log k\right)$
Vertex sim. (work):	$O\left(d^2\right)$	$O\left(\frac{B_X}{W}\right)$	$O\left(k\right)$
Vertex vim. (depth):	$O\left(\log d\right)$	$O\left(\log\left(\frac{B_X}{W}\right)\right)$	$O\left(\log k\right)$

TABLE VI: Advantages of ProbGraph in work and depth over exact baselines.

Application of ideas

$$\widehat{TC}_{\star} = \frac{1}{3} \sum_{(u,v)\in E} |\widehat{N_u \cap N_v}|_{\star}$$

where $\star$ indicates a specific $|\widehat{X \cap Y}|_{\star}$ estimator (cf. Table II).

Theorem VII.1. *Let $TC_{\star}$ be the estimator of the number of triangles. (cf. Section III). Then, depending on the underlying estimator $|\widehat{X \cap Y}|_{\star}$, we have the following cases:*

*For the **Bloom Filter** AND estimator, if $b\Delta \leq 0.499 B_X \log B_X$, then we have the following bound*

$$P\left(\left|TC - \widehat{TC}_{AND}\right| \geq t\right) \leq \frac{2\, m^2(1+o(1))\left(e^{\frac{\Delta b}{B_X - 1}} \frac{B_X}{b^2} - \frac{B_X}{b^2} - \frac{\Delta}{b}\right)}{9\, t^2}$$

*In the case of both **1-Hash** and k-**Hash** (below, we use the notation for 1-Hash), we have*

$$P\left(\left|TC - \widehat{TC}_{1H}\right| \geq t\right) \leq 2 \exp\left(-\frac{18\, k\, t^2}{\left(\sum_{v\in V} d(v)^2\right)^2}\right)$$

Moreover, if the maximum degree is Δ, then

$$P\left(\left|TC - \widehat{TC}_{1H}\right| \geq t\right) \leq 2 \exp\left(-\frac{9\, k\, t^2}{4\,(\Delta+1) \sum_{v\in V} d(v)^3}\right)$$

Application of ideas

- The k-Hash TC estimator is MLE
- Comparison to other triangle estimators like GAP, ASAP, MCMC
- Tested on datasets like SNAP, KONECT (K), DIMACS
- Tested on Dell PowerEdge R910 server with an Intel Xeon X7550 CPUs @ 2.00GHz with 18MB L3 cache, 1TiB RAM, and 32 cores per CPU

Experimental analysis

- Estimate size of intersection of neighborhoods (no one estimator performs best, but increasing storage space generally increases performance)

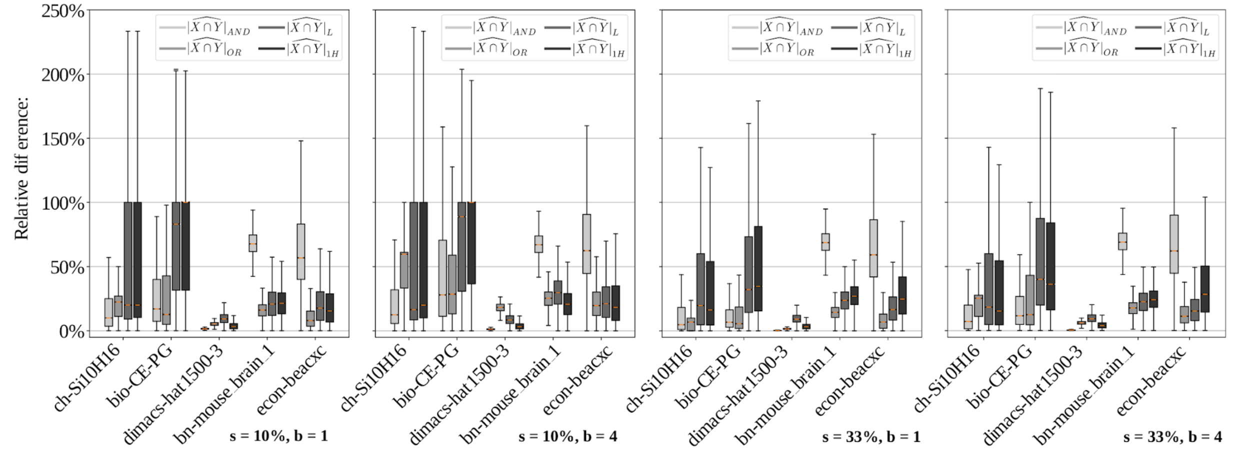


Fig. 3: Analysis of the accuracy of PG estimators of $|X \cap Y|$.

Experimental analysis

Min-Hash:

- Highest speedups
- lower memory requirements
- lower accuracy

Bloom Filter

- High accuracy
- High speedups

Experimental analysis

- Speedups of 30x or more over baselines while preserving 90% or higher accuracy
- Only about 25% extra storage needed

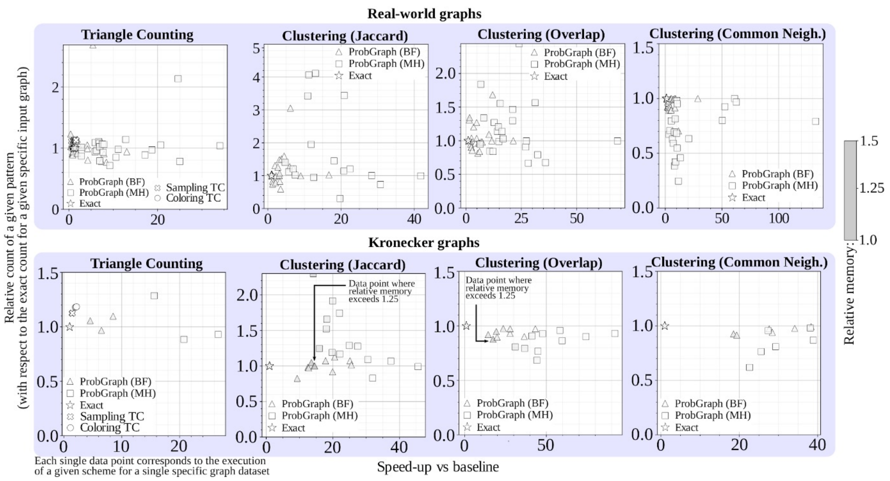


Fig. 4: Summary of advantages of PG for real-world (top panel) and Kronecker (bottom panel) graphs, for Triangle Counting and Clustering. All 32 cores are used. Note that most data points are white or almost white because they come with very low amounts of additional memory (we annotate a few data points that come with more than 25% additional relative memory amounts). Other graph problems come with the same performance/memory/accuracy patterns for the used comparison baselines.

Advantages of ProbGraph over previous work

- Theoretical bounds provided on estimators
- Observe better performance than previous “heuristic” methods
- Generalized approach to estimation, meaning these ideas can be applied uniformly to a wide range of problems
- Offers good (often strong) scaling because load balance issues mitigated by uniform data structure sizes
- Future work: try to develop or integrate other probabilistic set representations into ProbGraph
- Look for/analyze algorithms that require set unions instead of intersections (advantages offered to bloom filter)