Finding a Concise, Precise, and Exhaustive Set of Near Bi-Cliques in Dynamic Graphs (Supplementary Document)

A PROOF OF LEMMA 1

Proof. Let $\tilde{\mathcal{I}} = \tilde{\mathcal{S}} \cup \tilde{\mathcal{D}} \cup \tilde{\mathcal{T}}$ where $\tilde{\mathcal{S}} \subseteq \mathcal{S}$, $\tilde{\mathcal{D}} \subseteq \mathcal{D}$, and $\tilde{\mathcal{T}} \subseteq \mathcal{T}$; and $\tilde{\mathcal{I}}' = \tilde{\mathcal{S}}' \cup \tilde{\mathcal{D}}' \cup \tilde{\mathcal{T}}'$ where $\tilde{\mathcal{S}}' \subseteq \mathcal{S}$, $\tilde{\mathcal{D}}' \subseteq \mathcal{D}$, and $\tilde{\mathcal{T}}' \subseteq \mathcal{T}$. From the definition of $\mathcal{L}_{\mathcal{I}}(\mathcal{B}_{\tilde{\mathcal{I}}})$, Eq. (3) is equivalent to Eq. (14), which is equivalent to (15).

$$\frac{(|\tilde{\mathcal{S}}| + |\tilde{\mathcal{D}}| + |\tilde{\mathcal{T}}| + 1)}{|\tilde{\mathcal{S}}| \times |\tilde{\mathcal{D}}| \times |\tilde{\mathcal{T}}|} < \frac{(|\tilde{\mathcal{S}}'| + |\tilde{\mathcal{D}}'| + |\tilde{\mathcal{T}}'| + 1)}{|\tilde{\mathcal{S}}'| \times |\tilde{\mathcal{D}}'| \times |\tilde{\mathcal{T}}'|}. \tag{14}$$

$$\begin{aligned}&(|\tilde{\mathcal{S}}'| \times |\tilde{\mathcal{D}}'| \times |\tilde{\mathcal{T}}'|) \times (|\tilde{\mathcal{S}}| + |\tilde{\mathcal{D}}| + |\tilde{\mathcal{T}}| + 1)\\ &\qquad < (|\tilde{\mathcal{S}}| \times |\tilde{\mathcal{D}}| \times |\tilde{\mathcal{T}}|) \times (|\tilde{\mathcal{S}}'| + |\tilde{\mathcal{D}}'| + |\tilde{\mathcal{T}}'| + 1).\end{aligned} \tag{15}$$

Since $\mathcal{B}_{\tilde{\mathcal{I}}}$ is strictly larger than $\mathcal{B}_{\tilde{\mathcal{I}}'}$, $|\tilde{\mathcal{S}}| > |\tilde{\mathcal{S}}'|$, $|\tilde{\mathcal{D}}| > |\tilde{\mathcal{D}}'|$, or $|\tilde{\mathcal{T}}| > |\tilde{\mathcal{T}}'|$ holds. Without loss of generality, we assume that $|\tilde{\mathcal{S}}| > |\tilde{\mathcal{S}}'|$ holds. From $|\tilde{\mathcal{S}}| > |\tilde{\mathcal{S}}'| \geq 1$, $|\tilde{\mathcal{D}}| \geq |\tilde{\mathcal{D}}'| \geq 1$, and $|\tilde{\mathcal{T}}| \geq |\tilde{\mathcal{T}}'| \geq 1$, Eqs. (16)-(19) hold.

$$(|\tilde{\mathcal{S}}'| \times |\tilde{\mathcal{D}}'| \times |\tilde{\mathcal{T}}'|) < (|\tilde{\mathcal{S}}| \times |\tilde{\mathcal{D}}| \times |\tilde{\mathcal{T}}|). \tag{16}$$

$$(|\tilde{\mathcal{S}}'| \times |\tilde{\mathcal{D}}'| \times |\tilde{\mathcal{T}}'|) \times |\tilde{\mathcal{S}}| \leq (|\tilde{\mathcal{S}}| \times |\tilde{\mathcal{D}}| \times |\tilde{\mathcal{T}}|) \times |\tilde{\mathcal{S}}'|. \tag{17}$$

$$(|\tilde{\mathcal{S}}'| \times |\tilde{\mathcal{D}}'| \times |\tilde{\mathcal{T}}'|) \times |\tilde{\mathcal{D}}| < (|\tilde{\mathcal{S}}| \times |\tilde{\mathcal{D}}| \times |\tilde{\mathcal{T}}|) \times |\tilde{\mathcal{D}}'|. \tag{18}$$

$$(|\tilde{\mathcal{S}}'| \times |\tilde{\mathcal{D}}'| \times |\tilde{\mathcal{T}}'|) \times |\tilde{\mathcal{T}}| < (|\tilde{\mathcal{S}}| \times |\tilde{\mathcal{D}}| \times |\tilde{\mathcal{T}}|) \times |\tilde{\mathcal{T}}'|. \tag{19}$$

Eqs. (16)-(19) imply Eq. (15), which implies Eq. (3). □

B Q1. SEARCH QUALITY

In Figure 7, we divide the relative cost in Eq. (13) into three portions related to preciseness (i.e., $\phi_P(\mathcal{M})$ in Eq. (1)), exhaustiveness (i.e., $\phi_E(\mathcal{R})$ in Eq. (2)), and conciseness (i.e., $\phi_C(\mathcal{B})$ in (4)), respectively. The total cost was lowest in CutNPeel, and the near bi-cliques detected by it were especially superior in exhaustiveness and preciseness.

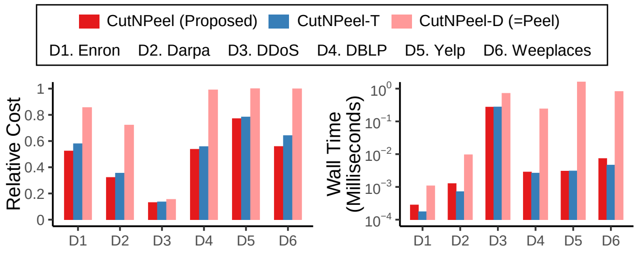
C Q3. APPLICATION: COMPRESSION

In Table 1, we report the compression rates of CutNPeel, Com2, and TimeCrunch with standard deviations. See the main paper for detailed experimental settings.

D Q4. ABLATION STUDY

In this section, we show the effectiveness of adaptive thresholds and partitioning in CutNPeel. We consider two variants:

- **CutNPeel-T**: CutNPeel with fixed thresholds ($\theta(t) = 0$).



(a) Relative Cost (b) Wall Time per Near Bi-clique

Figure 6: Adaptive thresholds and partitioning in CutNPeel improve the quality of detected near bi-cliques. Partitioning also reduces time taken for detecting each near bi-clique.

- **CutNPeel-D (=Peel)**: CutNPeel without partitioning.

As seen in Figure 6a, both adaptive thresholds and partitioning consistently improved the quality of detected near bi-cliques, and partitioning also reduced time taken per near bi-clique (see Section 4.2.1) of the main paper for further discussion).

E Q5. PARAMETER ANALYSIS

In this section, we examine the effects of the parameters of CutNPeel. Specifically, we measured how the threshold decrement rate α and the number of iterations T affect the running time and relative cost in Eq. (13) of CutNPeel.

As seen in Figure 8, increasing α (i.e., slowing down the decrease of $\theta(t)$ for more exploration) improved the quality of near bi-cliques detected by CutNPeel while increasing the running time.

Similarly, the quality of near bi-cliques detected by CutNPeel got better as we increased the number of iterations T, as seen in Figure 9. Diminishing returns were apparent. That is, the improvement due to an additional iteration decreased as T increased, and the quality almost converged within 80 iterations. Especially, on the DBLP, Yelp, and Weeplaces datasets, CutNPeel terminated within 40 iterations, and thus the running time did not increase when T was over 40.

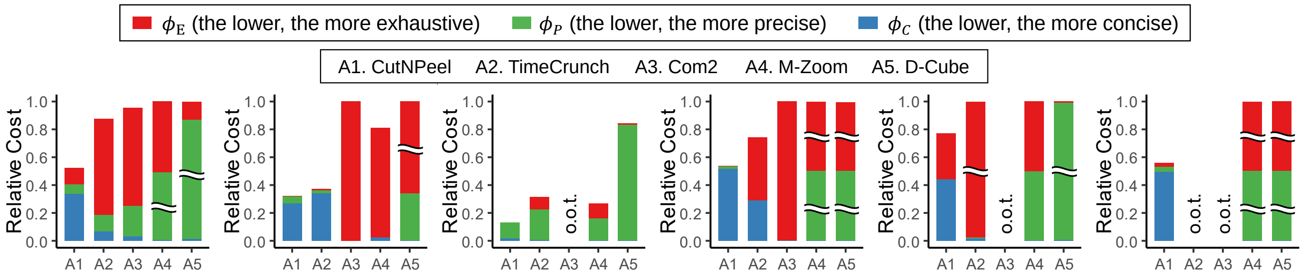
F Q6. APPLICATION: PATTERN DISCOVERY

In this section, we introduce some interesting patterns that CutNPeel detects as near bi-cliques on the DBLP and DDoS datasets.

<u>**DBLP:**</u> A near bi-clique of density (i.e., the ratio of missinge edges) 0.771 that is visualized in Figure 1c in Section 1 reveals 7 researchers who presented at the 4 same venues for 10 consecutive years from 1995 to 2004 with few exceptions. Noticeably, the venues, which are International Test Conference, Asian Test Symposium, IEEE VLSI Test Symposium, and VLSI Design Conference are on similar topics. Another near bi-clique of density 0.875 reveals 8 researchers who presented at European Conference on Advances in Databases and Information Systems in 5 non-consecutive years with few exceptions. See Table 2 for the lists of the researchers.

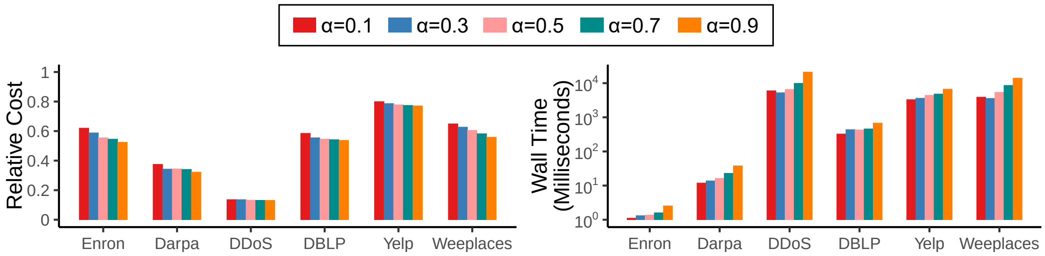
Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456



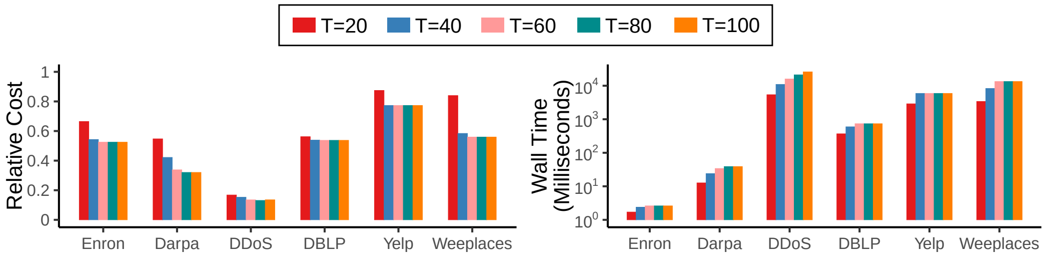
(a) Enron (b) Darpa (c) DDoS (d) DBLP (e) Yelp (f) Weeplaces

Figure 7: CutNPeel provides near bi-cliques with the best quality, and they are especially superior in exhaustiveness and preciseness. In each plot, we divide the relative cost in Eq. (13) (the lower the relative cost is, the better the quality of near bi-cliques is) into three portions that are related to preciseness, exhaustiveness, and conciseness, respectively.



(a) Relative Cost Depending on α (b) Running Time Depending on α

Figure 8: Effects of α. As the threshold decrement rate α increases, the quality of near bi-cliques detected by CutNPeel gets better, at the expense of speed.



(a) Relative Cost Depending on T (b) Running Time Depending on T

Figure 9: Effects of T. As the number of iterations T increases, the quality of near bi-cliques detected by CutNPeel gets better with diminishing returns. The quality almost converges within 80 iterations.

DDoS: On the DDoS dataset, many near bi-cliques detected by CutNPeel were composed by one source IP and multiple destination IPs and those composed by multiple source IPs and one destination IP. For example, a near bi-clique consists of one source IP and 5,021 destination IPs between which 226,802 connections were made within during 51 non-consecutive seconds. The bi-clique reveals networks attacks at around 21:13:00 UTC on August, 4, 2007.

REFERENCES

[1] Miguel Araujo, Spiros Papadimitriou, Stephan Günnemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, Evangelos E Papalexakis, and Danai Koutra. 2014. Com2: fast automatic discovery of temporal ('comet') communities. In *PAKDD*.

[2] Neil Shah, Danai Koutra, Tianmin Zou, Brian Gallagher, and Christos Faloutsos. 2015. Timecrunch: Interpretable dynamic graph summarization. In *KDD*.

Table 1: CutNPeel is effective in lossless compression of dynamic graphs. It consistently achieved the best compression. o.o.t.: out of time (> 6 hours).

Metric*	Method	**Compression Rates in %** (the lower, the better)					
		Enron	Darpa	DDoS	DBLP	Yelp	Weeplaces
Eq. (13)	CutNPeel	52.45±0.40	32.27±0.66	13.12±0.58	53.73±0.081	**77.11±0.18**	55.89±0.31
	TimeCrunch	87.7±0.0	37.2±0.0	31.5±0.0	74.2±0.0	100.6±0.0	o.o.t.
	Com2	95.52±1.56	100±0.0	o.o.t.	99.98±0.035	o.o.t.	o.o.t.
[2]**	TimeCrunch	86.3±0.0	36.7±0.0	16.3±0.0	79.7±0.0	100.0±0.0	o.o.t.
[1]	CutNPeel	**48.51±0.56**	**27.58±0.86**	**8.0**	**52.53±0.36**	77.77±0.25	**52.38±0.062**
	TimeCrunch	79.3±0.0	34.6±0.0	22.8±0.0	63.9±0.0	100.2±0.0	o.o.t.
	Com2	59.17±0.74	202.98±0.0	o.o.t.	57.51±0.034	o.o.t.	o.o.t.

* Different metrics are based on different encoding methods.

** The encoding method used in [2] is not applicable to CutNPeel and Com2.

Table 2: Example patterns captured as near bi-cliques by CutNPeel on the DBLP dataset. The first one reveals seven researchers who presented at the four same venues for ten consecutive years with few exceptions. Note that the four conferences are on relevant topics.

Author	Venue	Year	Number of Objects	Density
{Nur A. Touba, Yervant Zorian, Sudhakar M. Reddy, Irith Pomeranz, Jacob A. Abraham, Vishwani D. Agrawal, A. J. van de Goor}	{International Test Conference, Asian Test Symposium, IEEE VLSI Test Symposium, VLSI Design Conference}	{1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004}	7/4/10	0.771
{Yannis Manolopoulos, Jaroslav Pokorn, Kjetil Nrvg, Tadeusz Morzy, Marek Wojciechowski, Maciej Zakrzewicz, Alexey Tsymbal, Leonid A. Kalinichenko}	{European Conference on Advances in Databases and Information Systems}	{1997, 1998, 1999, 2001, 2003}	8/1/5	0.875