CoreScope: Graph Mining Using k-Core Analysis - Patterns, Anomalies and Algorithms (Supplementary Document)

Kijung Shin
Carnegie Mellon University
Pittsburgh, PA, USA
kijungs@cs.cmu.edu

Tina Eliassi-Rad
Northeastern University
Boston, MA, USA
eliassi@ccs.neu.edu

Christos Faloutsos
Carnegie Mellon University
Pittsburgh, PA, USA
christos@cs.cmu.edu

***Abstract*—In this supplementary document, we provide additional proofs and experimental results, which supplement the main paper [1].**

I. Proofs

A. *Proof of Lemma 1*

In this section, we prove Lemma 1 in the main paper. For the proof, we use Lemmas 3 and 4, which give upper and lower bounds of degeneracy.

Lemma 3 (Lower Bound of Degeneracy [2])**.** *The half of the average degree lower bounds the degeneracy. Let d_{avg} be the average degree. Then, $k_{max} \geq \lceil m/n \rceil \geq d_{avg}/2$.*

Lemma 4 (Upper Bound of Degeneracy)**.** *The largest eigenvalue upper bounds the degeneracy. Let λ_1 be the largest eigenvalue of the adjacency matrix. Then $k_{max} \leq \lambda_1$.*

Proof. Let H be the degeneracy-core (i.e., k_{max}-core) of G and $d_{min}(H)$ be its minimum degree. By the definition of the k-core and degeneracy, $d_{min}(H) = k_{max}(G)$. Since the largest eigenvalue is lower bounded by minimum degree [3], $k_{max}(G) = d_{min}(H) \leq \lambda_1(H)$. The largest eigenvalue of a graph is also lower bounded by that of its induced subgraph [3]. Since the degeneracy-core is an induced subgraph due to its maximality, $k_{max}(G) \leq \lambda_1(H) \leq \lambda_1(G) = \lambda_1$. ∎

Lemma 5 states that the graph measures used for upper and lower bounding degeneracy in Lemmas 3 and Lemma 4 increase exponentially with q, the power of Kronecker products, in Kronecker Model.

Lemma 5. (Graph Measures Increasing Exponentially in Kronecker Graphs). *The average degree, the degeneracy, and the largest eigenvalue increase exponentially with q in $\{C_q\}_{q \geq 1}$, graphs generated by Kronecker Model.*
(1) $d_{avg}(G_q) = (d_{avg}(G_1))^q$, $\forall q \geq 1$.
(2) $k_{max}(G_q) \geq (k_{max}(G_1))^q$, $\forall q \geq 1$.
(3) $\lambda_1(G_q) = (\lambda_1(G_1))^q$, $\forall q \geq 1$.

Proof. Let $n(G)$ be the number of vertices and $nz(G)$ be the number of non-zero entries in the adjacency matrix. Then, $d_{avg}(G) = nz(G)/n(G)$. As $n(G_q) = (n(G_1))^q$ and $nz(G_q) = (nz(G_1))^q$, $d_{avg}(G_q) = nz(G_q)/n(G_q) = (nz(G_1))^q/(n(G_1))^q = (nz(G_1)/n(G_1))^q = (d_{avg}(G_1))^q$, $\forall q \geq 1$.

For seed graph G_1, $k_{max}(G_1) \geq (k_{max}(G_1))^1$. Assume $k_{max}(G_i) \geq (k_{max}(G_1))^i$. Each vertex in G_{i+1} can be represented as an ordered pair (v_i, v_1) where v_i is a vertex of G_i and v_1 is a vertex of G_1. Two vertices, (v_i, v_1) and (v'_i, v'_1), in G_{i+1} are adjacent if and only if v_i and v'_i are adjacent in G_i and v_1 and v'_1 are adjacent in G_1 [4]. Let $G'_i(V'_i, E'_i)$ be the degeneracy-core of $G_i(V_i, E_i)$ where $V'_i = \{v_i \in V_i | c(v_i) = k_{max}(G_i)\}$. Then, each vertex (v_i, v_1) in $S = \{(v_i, v_1) \in V_{i+1} | v_i \in V'_i, v_1 \in V'_1\}$ are adjacent to $d_{G'_i}(v_i) \times d_{G'_1}(v_1) (\geq k_{max}(G_i) \times k_{max}(G_1))$ vertices in S. Therefore, $k_{max}(G_{i+1}) \geq k_{max}(G_i) \times k_{max}(G_1) \geq k_{max}(G_1)^{(i+1)}$. By induction, $k_{max}(G_q) \geq (k_{max}(G_1))^q$, $\forall q \geq 1$.

Let $\lambda(G) = (\lambda_1, ..., \lambda_n)$ be the eigenvalues of the adjacency matrix of G, and $\lambda_1(G)$ be the largest eigenvalue. Then, $\lambda(G_q) = sort(\lambda(G_{q-1}) \otimes \lambda(G_1))$ [5]. As $\lambda_1(G_q) = \lambda_1(G_{q-1}) \times \lambda_1(G_1)$, $\lambda_1(G_q) = (\lambda_1(G_1))^q$, $\forall q \geq 1$. ∎

Proof of Lemma 1

Proof. Lemma 1 is proved by Lemmas 3, 4, and 5. ∎

B. *Proof of Lemma 2*

In this section, we prove Lemma 2 in the main paper. For the proof, we have to deal with self-loops in Kronecker graphs which happen naturally. We add one to the degree for each self-loop and define *a triangle in Kronecker graphs* as an unordered vertex triplet, which can contain multiple instances of the same vertex, where every instance is connected to all others either by self-loops or other edges. For example, (v_1, v_1, v_2) is a triangle in Kronecker graphs if v_1 has a self-loop and v_1 and v_2 are adjacent. Note that Lemma 2 and Theorem 1 (in the main paper) hold equally, with the original definitions of degree and a triangle, in Kronecker graphs without self-loops.

Proof of Lemma 2

Proof. Let $\lambda(G_i) = (\lambda_1(G_i), ..., \lambda_{n^i}(G_i))$ be the eigenvalues of the adjacency matrix of G_i. The number of walks of length 3 in G_i that begin and end on the same vertex is $\sum_{j=1}^{n^i}(\lambda_j(G_i))^3$ [6] and linearly related to the number of triangles, i.e., $\#\Delta(G_i) = \Theta(\sum_{j=1}^{n^i}(\lambda_j(G_i))^3)$. For seed graph G_1, $\sum_{j=1}^{n}(\lambda_j(G_1))^3 = (\sum_{j=1}^{n}(\lambda_j(G_1))^3)^1$. Assume $\sum_{j=1}^{n^i}(\lambda_j(G_i))^3 = (\sum_{j=1}^{n}(\lambda_j(G_1))^3)^i$. As $\lambda(G_{i+1}) = sort(\lambda(G_i) \otimes \lambda(G_1))$ [5],

$$\sum_{j=1}^{n^{(i+1)}} (\lambda_j(G_{i+1}))^3 = \sum_{r=1}^{n^i}\sum_{s=1}^{n}(\lambda_r(G_i))^3(\lambda_s(G_1))^3$$
$$= (\sum_{r=1}^{n^i}(\lambda_r(G_i))^3)(\sum_{s=1}^{n}(\lambda_s(G_1))^3) = \left(\sum_{s=1}^{n}(\lambda_s(G_1))^3\right)^{(i+1)}.$$

By induction, $\sum_{j=1}^{n^q}(\lambda_j(G_q))^3 = (\sum_{j=1}^{n}(\lambda_j(G_1))^3)^q$, $\forall q \geq 1$. Hence, $\#\Delta(G_q) = \Theta(\sum_{j=1}^{n^q}(\lambda_j(G_q))^3) = \Theta((\sum_{j=1}^{n}(\lambda_j(G_1))^3)^q)$, $\forall q \geq 1$. ■

C. *Proof of Theorem 2*

In this section, we prove Theorem 2 in the main paper.

Proof. From $p = \Omega(\log n/n)$, there exists $c > 0$ such that $p \geq c \log n/n$. Let $\epsilon = \max(2, 12/c)$ (> 1). Then,

$$\begin{aligned}
&P(\exists v \in V \text{ s.t. } d(v) > (1+\epsilon)(n-1)p) \\
&\quad \leq nP(d(v) > (1+\epsilon)(n-1)p) && \text{(Boole's inequality)} \\
&\quad \leq n\exp\{-(n-1)p\epsilon/3\} && \text{(Chernoff bound)} \\
&\quad \leq n\exp\{-c\log(n)(n-1)\epsilon/3n\} && (p \geq c\log n/n) \\
&\quad \leq n\exp\{-4\log(n)(n-1)/n\} && (\epsilon \geq 12/c) \\
&\quad \leq n\exp\{-2\log n\} = n^{-1}.
\end{aligned}$$

Let $q = P(\exists v \in V \text{ s.t. } d(v) > (1+\epsilon)(n-1)p)$. Then,

$$\begin{aligned}
E[k_{max}] &\leq E[d_{max}] \leq (1-q)(1+\epsilon)(n-1)p + q(n-1) \\
&\leq (1+\epsilon)(n-1)p + (n-1)/n = O(np)
\end{aligned}$$

Hence, $E[k_{max}] = O(np)$. As $E[k_{max}] \geq E[d_{avg}/2] = \Omega(np)$ by Lemma 3, $E[k_{max}] = \Theta(np)$.

On the other hand, the expected number of triangles is the sum of probabilities that each three vertices form a traingle:

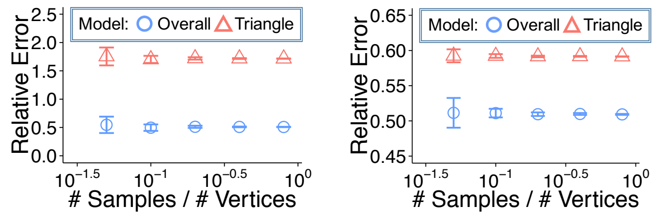
$$E[\#\Delta] = \frac{n(n-1)(n-2)}{6}p^3.$$

Therefore, $E[\#\Delta] = \Theta(n^3p^3) = \Theta(E[k_{max}]^3)$. ■

II. Additional Experiments

A. Core-D *with Smaller Number of Samples*

Figure 1 presents the accuracy of Core-D with different sample sizes in the two largest datasets. Even with small number of samples less than the number of vertices, Core-D, especially Overall Model, accurately and reliably estimated degeneracy. Thus, Core-D is still effective even when the amount of available memory space is less than n.



(a) Friendster (b) Twitter

Fig. 1: **Core-D is nimble and accurate.** Points and error bars represent the average accuracy and ± one standard deviation over ten runs, respectively. Core-D reliably estimates degeneracy even with small number of samples less than the number of vertices.

B. Core-S *with Different Numbers of Spreaders*

In the main paper, we compared the average influence of the ten vertices chosen by Core-S with that of the vertices chosen by other influential spreader identification methods. In this section, we compared the methods when different numbers of spreaders are chosen. Specifically, for different k values, we compared the average influence of k vertices chosen by Core-S with that of the vertices chosen by the following methods:

- K-Core [7]: all vertices with the highest coreness.
- K-Truss [8]: all vertices with the highest truss number.
- Eigenvector Centrality (EC) [9]: top-k vertices with the highest eigenvector centralities in the entire graph.

As in the main paper, we measured the influence of each vertex using SIR simulation (see Appendix B in the main paper for details) and also compared the time taken for choosing influential vertices in each method.

Figure 2 presents the results in social networks, where influential spreader identification has been used. Regardless of k, Core-S provided the best trade-off between speed and accuracy. Specifically, the average influence of the vertices chosen by Core-S was up to **2.6× higher** than that of all the vertices in the degeneracy-core (K-Core) although the gap decreases as k increases. However, additional time taken in Core-A for further refining vertices in degeneracy-cores was at most 12% of the time taken for the core decomposition of entire graphs. Besides, Core-S was up to **17× faster**, than EC, which has to compute the eigenvector centrality in entire graphs (instead of only in degeneracy-cores). However, the average influence of the vertices chosen by Core-S was comparable with that of the vertices found by EC (100% in Orkut, 97-104% in Flickr, 99-100% in Catster, 88-100% in Youtube, and 95-100% in Email).

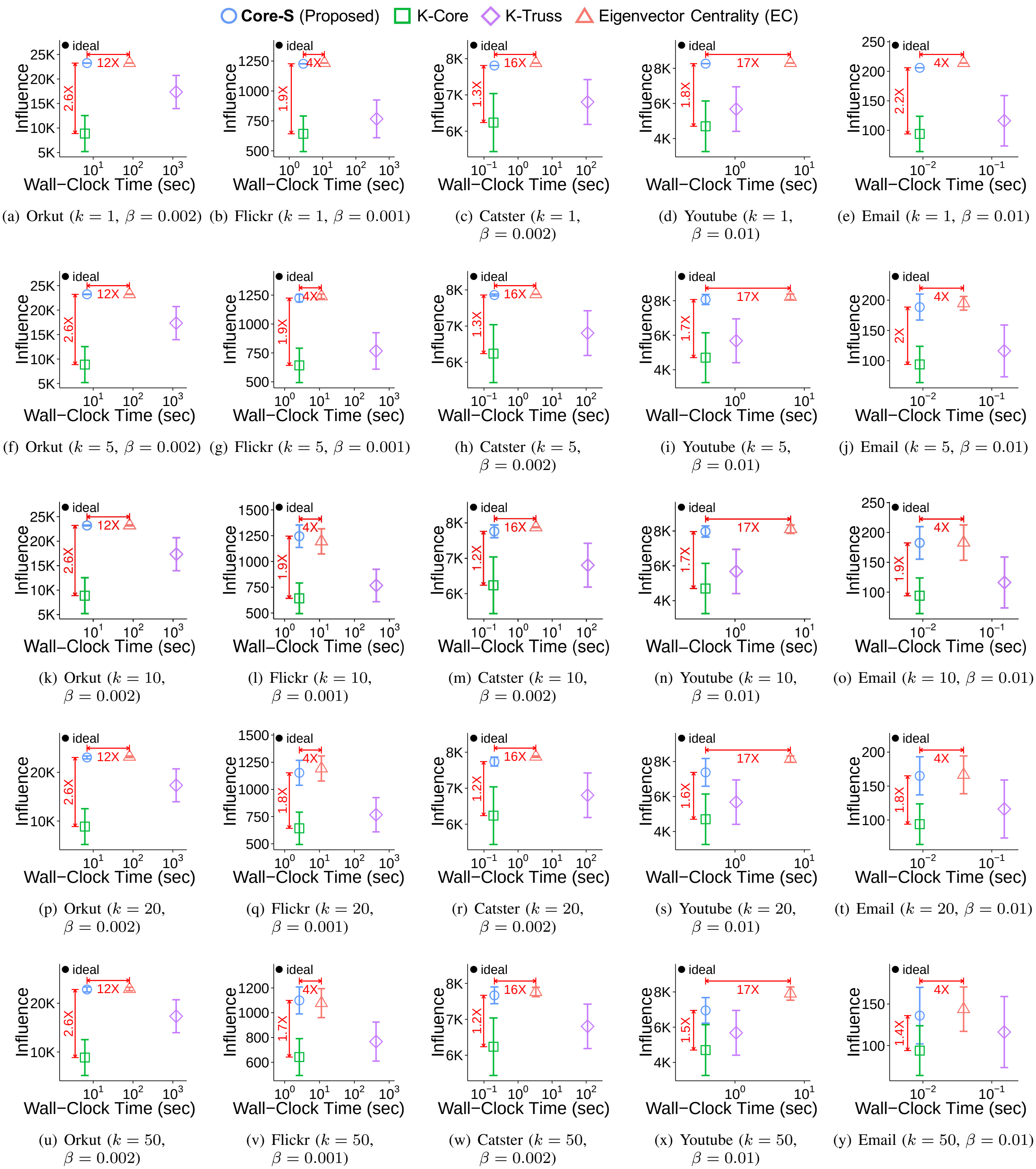
References

[1] K. Shin, T. Eliassi-Rad, and C. Faloutsos, "Corescope: Graph mining using k-core analysis - patterns, anomalies and algorithms," in *ICDM*, 2016.

[2] P. Erdös, "On the structure of linear graphs," *Israel J. of Math.*, vol. 1, no. 3, pp. 156–160, 1963.

[3] A. E. Brouwer and W. H. Haemers, *Spectra of graphs*. Springer Science & Business Media, 2011.

[4] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos, "Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication," in *PKDD*, 2005, pp. 133–145.



(a) Orkut ($k = 1$, $\beta = 0.002$) (b) Flickr ($k = 1$, $\beta = 0.001$) (c) Catster ($k = 1$, $\beta = 0.002$) (d) Youtube ($k = 1$, $\beta = 0.01$) (e) Email ($k = 1$, $\beta = 0.01$)

(f) Orkut ($k = 5$, $\beta = 0.002$) (g) Flickr ($k = 5$, $\beta = 0.001$) (h) Catster ($k = 5$, $\beta = 0.002$) (i) Youtube ($k = 5$, $\beta = 0.01$) (j) Email ($k = 5$, $\beta = 0.01$)

(k) Orkut ($k = 10$, $\beta = 0.002$) (l) Flickr ($k = 10$, $\beta = 0.001$) (m) Catster ($k = 10$, $\beta = 0.002$) (n) Youtube ($k = 10$, $\beta = 0.01$) (o) Email ($k = 10$, $\beta = 0.01$)

(p) Orkut ($k = 20$, $\beta = 0.002$) (q) Flickr ($k = 20$, $\beta = 0.001$) (r) Catster ($k = 20$, $\beta = 0.002$) (s) Youtube ($k = 20$, $\beta = 0.01$) (t) Email ($k = 20$, $\beta = 0.01$)

(u) Orkut ($k = 50$, $\beta = 0.002$) (v) Flickr ($k = 50$, $\beta = 0.001$) (w) Catster ($k = 50$, $\beta = 0.002$) (x) Youtube ($k = 50$, $\beta = 0.01$) (y) Email ($k = 50$, $\beta = 0.01$)

Fig. 2: **CORE-S achieves both speed and accuracy.** β denotes the infection rate in SIR Model. Points in each plot represent the performances of different methods. Upper-left region indicates better performance. CORE-S provided the best trade-off between speed and accuracy. Specifically, it found up to **2.6× more influential** vertices than K-CORE with similar speed. Compared with EC, CORE-S was up to **17× faster**, while still finding vertices with comparable influence (100% in Orkut, 97-104% in Flickr, 99-100% in Catster, 88-100% in Youtube, and 95-100% in Email).

[5] C. F. Van Loan, “The ubiquitous kronecker product,” *J. of comp. and appl. math.*, vol. 123, no. 1, pp. 85–100, 2000.

[6] C. E. Tsourakakis, “Fast counting of triangles in large real networks without counting: Algorithms and laws,” in *ICDM*, 2008.

[7] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, “Identification of influential spreaders in complex

networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[8] M.-E. G. Rossi, F. D. Malliaros, and M. Vazirgiannis, "Spread it good, spread it fast: Identification of influential nodes in social networks," in *World Wide Web Companion*, 2015.

[9] B. Macdonald, P. Shakarian, N. Howard, and G. Moores, "Spreaders in the network sir model: An empirical study," *arXiv preprint arXiv:1208.4269*, 2012.