SPECGREEDY

Unified Dense Subgraph Detection

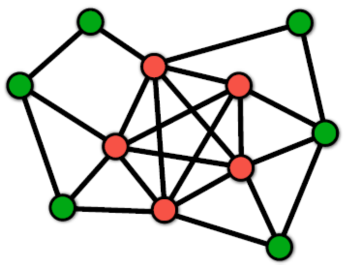
Wenjie Feng[+], Shenhua Liu[+], Danai Koutra[#],
Huawei Shen[+], Xueqi Cheng[+]

[+]*Institute of Computing Technology, ICT, CAS*

[#]*University of Michigan, Ann Arbor*

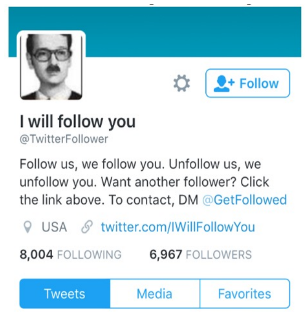
Motivation: Dense Subgraph Detection

- Ubiquitous applications



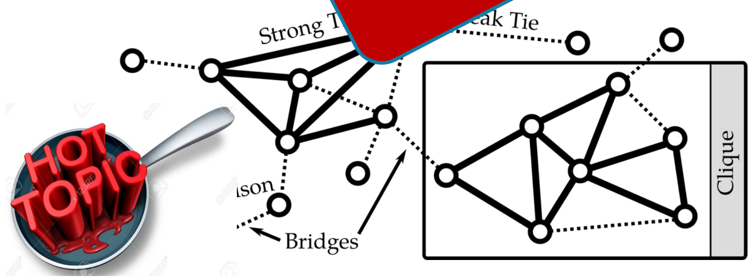
Densest subgraph
[Charikar M.'0

detection
uchi A. et al'18]

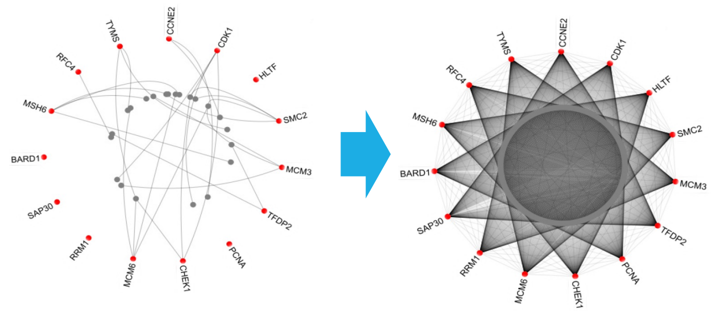


Fraud detection
[B. Hooi et al'18]

unified formalization ?



News hot topics discovery
[LY, Chu et al'12]

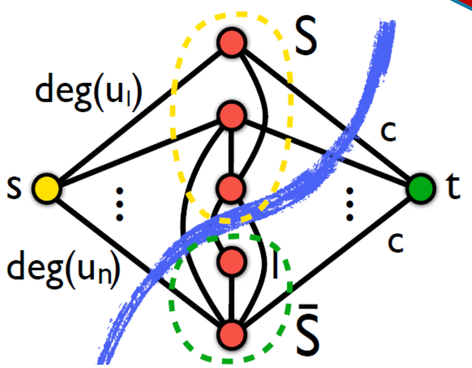


Contrast dense subgraph
[Wong S. W e al'18]

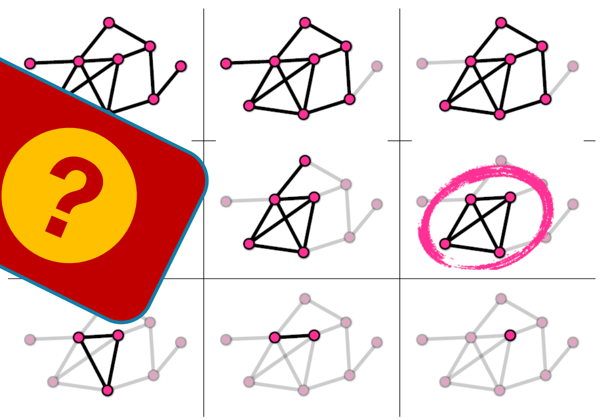
Motivation: Densest Subgraph Detection

- **Polynomial-time methods**
  - **Exact solution**: maximum flows based algo.
    - $O(|U|^2 + |U| \cdot |E|)$
  - **Approx. solution**: Greedy Peeling algo. ($^1/_2$-optimal)
    - $O(|E| \cdot \log|U|)$

faster algorithm ?



Max-flow exact DS
[Goldberg’84]



Greedy approx. DS
[Charikar,M.’00]

OVERVIEW-METHOD

- GENDS: a unified problem framework

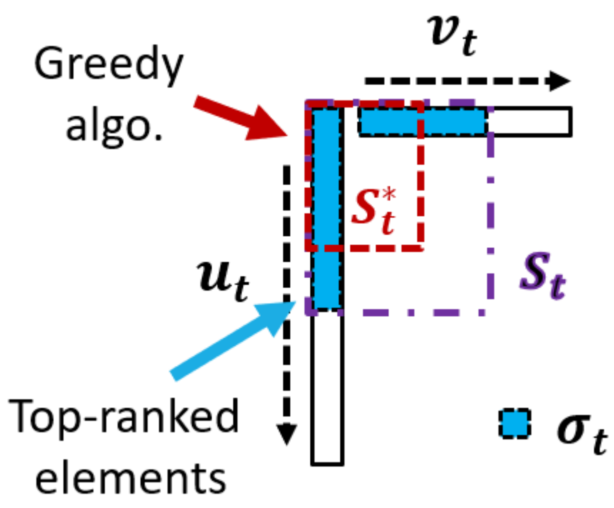
$$S^* = \arg\max_{S \subseteq V, |S| \geq 1} g(S; \mathbf{P}, \mathbf{Q}) = \arg\max_{\boldsymbol{x} \in \{0,1\}^n, |\boldsymbol{x}| \geq 1} \frac{\boldsymbol{x}^T \mathbf{P} \boldsymbol{x}}{\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}}$$

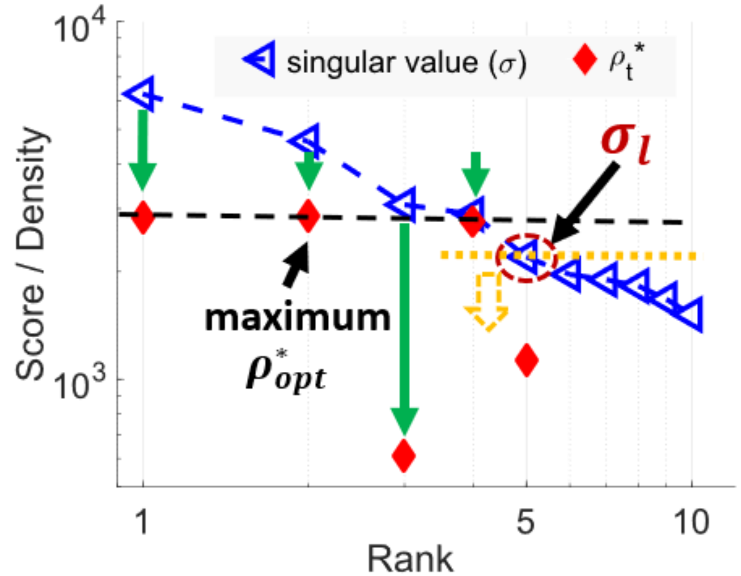
$$\mathbf{P} = \mathbf{A} + 2\mathbf{D}_{\boldsymbol{c}} \qquad \mathbf{Q} = \mathbf{A}' + \gamma \mathbf{I}$$

MinQuotientCut [Chung RK. Fan'96]	***SparseCutDS*** [Miyauchi A. et al'18]	***TempDS*** [Wong S. W e al'18]
Fraudar [B. Hooi et al'18]	***Charikar*** [Charikar M.'00]	***Risk-averse DS*** [C. E. Tsourakakis et al'19]

OVERVIEW-METHOD

- **SPECGREEDY**: a unified detection algorithm

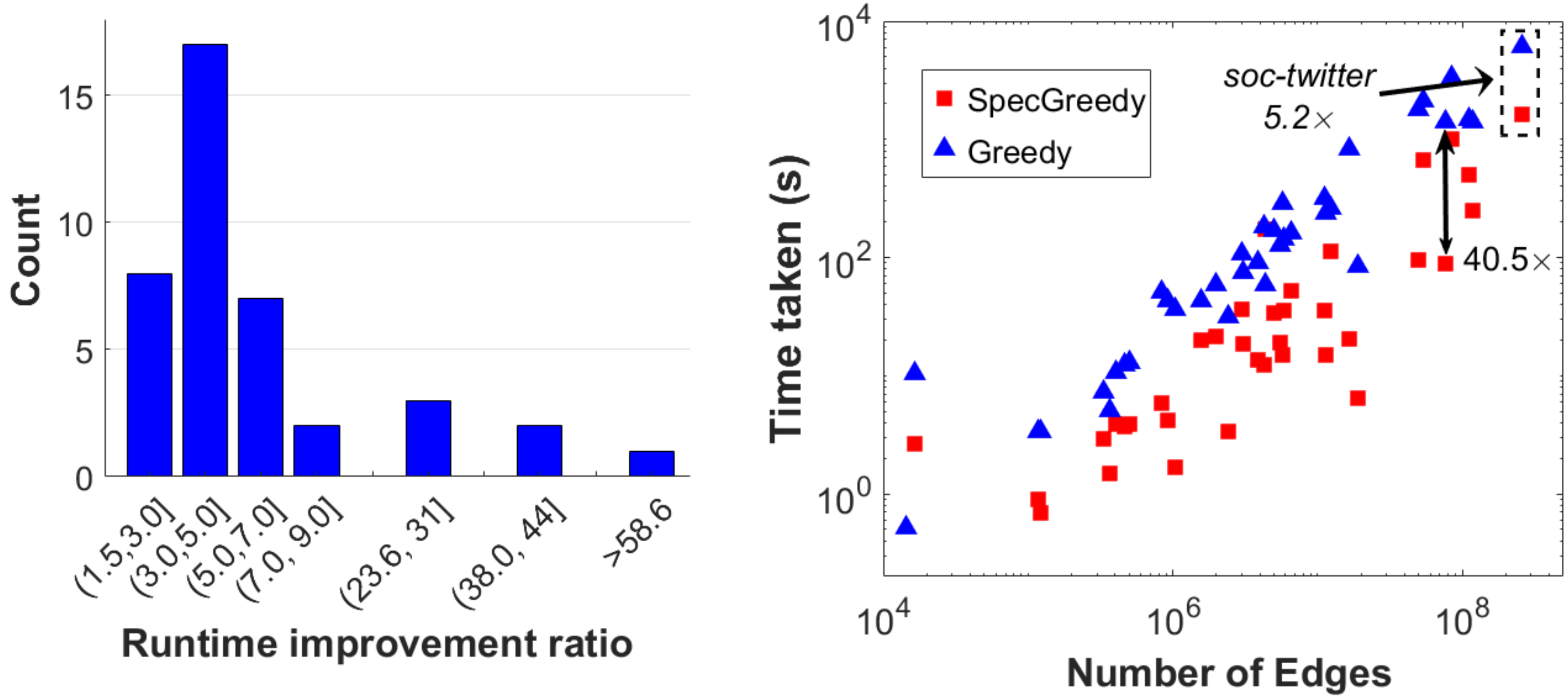




Graph spectral + Greedy approximation

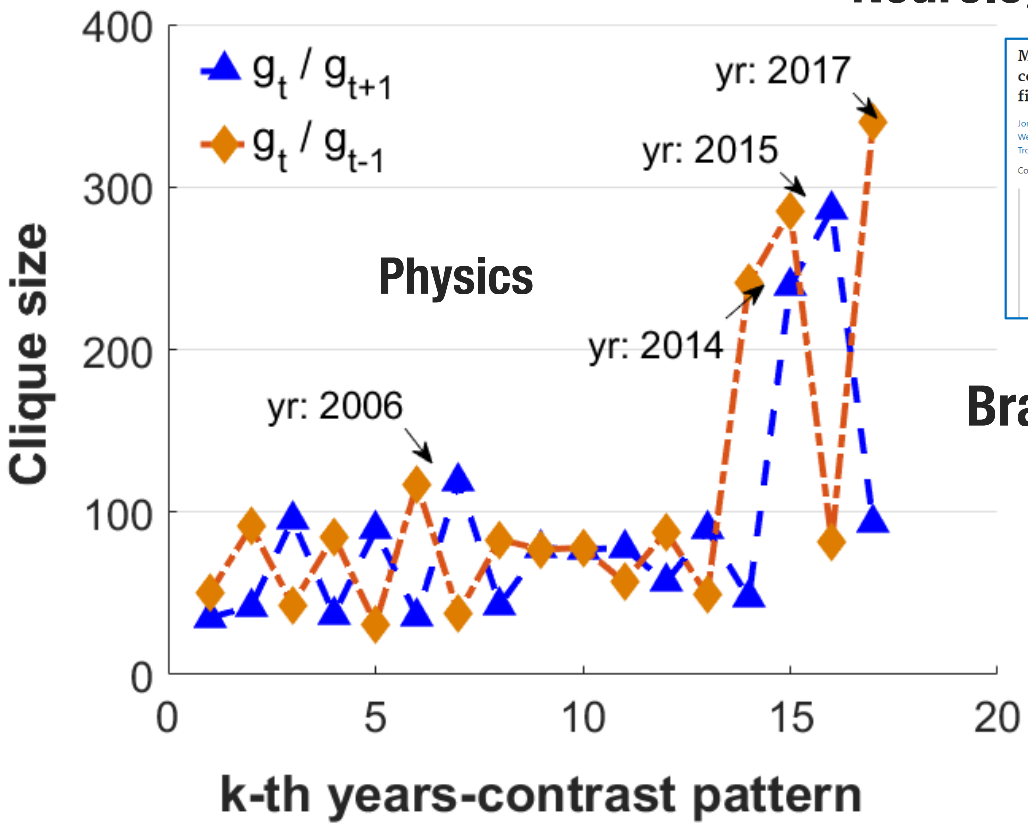
Linearly scalable with the graph size

OVERVIEW-PERFORMANCE

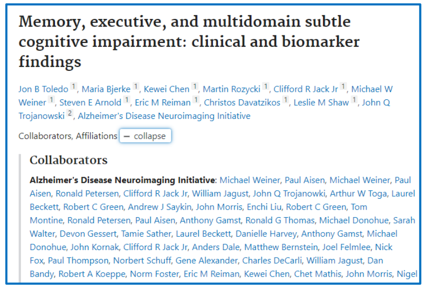


SPECGREEDY is faster than Greedy algo.

OVERVIEW-PERFORMANCE



Neurology & Medicine



Memory, executive, and multidomain subtle cognitive impairment: clinical and biomarker findings

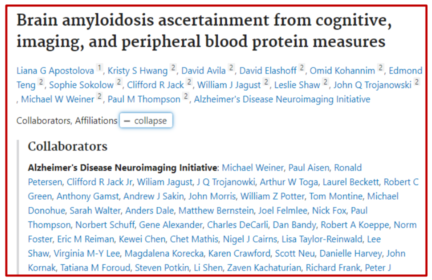
Jon B Toledo [1], Maria Bjerke [1], Kewei Chen [1], Martin Rozycki [1], Clifford R Jack Jr [1], Michael W Weiner [1], Steven E Arnold [1], Eric M Reiman [1], Christos Davatzikos [1], Leslie M Shaw [1], John Q Trojanowski [2], Alzheimer's Disease Neuroimaging Initiative

Collaborators, Affiliations – collapse

Collaborators

Alzheimer's Disease Neuroimaging Initiative: Michael Weiner, Paul Aisen, Michael Weiner, Paul Aisen, Ronald Petersen, Clifford R Jack Jr, William Jagust, John Q Trojanowki, Arthur W Toga, Laurel Beckett, Robert C Green, Andrew J Saykin, John Morris, Enchi Liu, Robert C Green, Tom Montine, Ronald Petersen, Paul Aisen, Anthony Gamst, Ronald G Thomas, Michael Donohue, Sarah Walter, Devon Gessert, Tamie Sather, Laurel Beckett, Danielle Harvey, Anthony Gamst, Michael Donohue, John Kornak, Clifford R Jack Jr, Anders Dale, Matthew Bernstein, Joel Felmlee, Nick Fox, Paul Thompson, Norbert Schuff, Gene Alexander, Charles DeCarli, William Jagust, Dan Bandy, Robert A Koeppe, Norm Foster, Eric M Reiman, Kewei Chen, Chet Mathis, John Morris, Nigel

Brain network and Disease



Brain amyloidosis ascertainment from cognitive, imaging, and peripheral blood protein measures

Liana G Apostolova [1], Kristy S Hwang [2], David Avila [2], David Elashoff [2], Omid Kohannim [2], Edmond Teng [2], Sophie Sokolow [2], Clifford R Jack [2], William J Jagust [2], Leslie Shaw [2], John Q Trojanowski [2], Michael W Weiner [2], Paul M Thompson [2], Alzheimer's Disease Neuroimaging Initiative

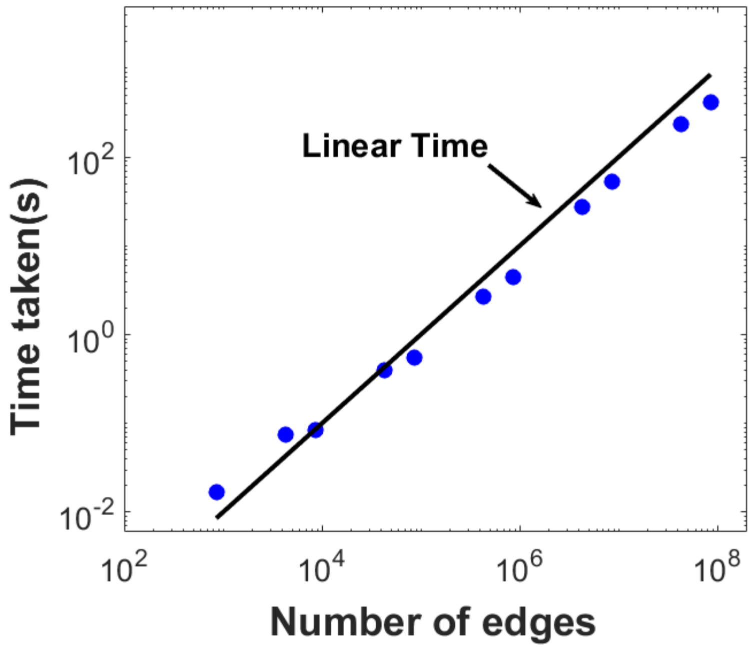
Collaborators, Affiliations – collapse

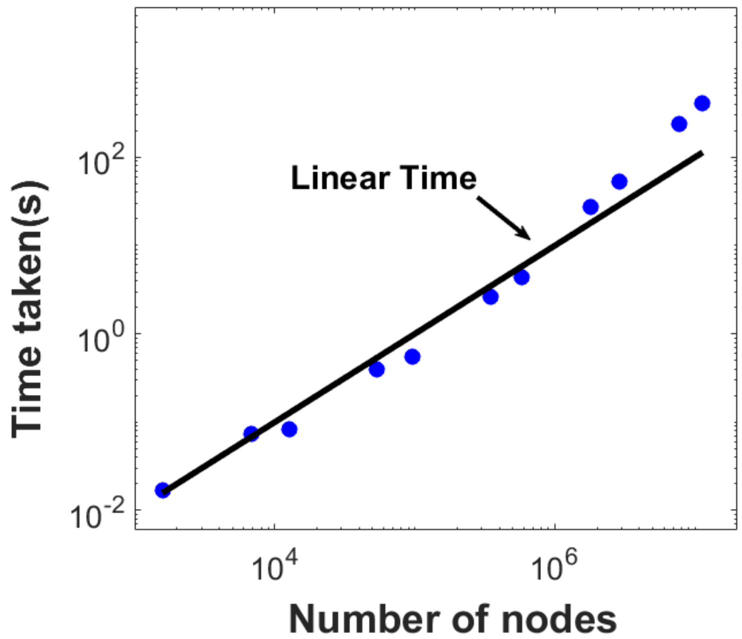
Collaborators

Alzheimer's Disease Neuroimaging Initiative: Michael Weiner, Paul Aisen, Ronald Petersen, Clifford R Jack Jr, Wiliam Jagust, J Q Trojanowki, Arthur W Toga, Laurel Beckett, Robert C Green, Anthony Gamst, Andrew J Sakin, John Morris, William Z Potter, Tom Montine, Michael Donohue, Sarah Walter, Anders Dale, Matthew Bernstein, Joel Felmlee, Nick Fox, Paul Thompson, Norbert Schuff, Gene Alexander, Charles DeCarli, Dan Bandy, Robert A Koeppe, Norm Foster, Eric M Reiman, Kewei Chen, Chet Mathis, Nigel J Cairns, Lisa Taylor-Reinwald, Lee Shaw, Virginia M-Y Lee, Magdalena Korecka, Karen Crawford, Scott Neu, Danielle Harvey, John Kornak, Tatiana M Foroud, Steven Potkin, Li Shen, Zaven Kachaturian, Richard Frank, Peter J

***SpecGreedy* is effective**, *e.g. contrast dense subgraphs*

OVERVIEW-PERFORMANCE





SpecGreedy is scalable

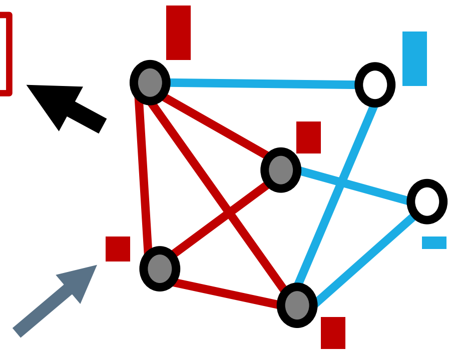
Road Map

- Introduction
- Proposed Method:
  - Unified Formulation: **GENDS** <<
  - Algorithm: SPECGREEDY
- Experiments
- Conclusion

GENDS: Unified Formulation

- GENDS: Generalized Dense Subgraph detection
- **Given:**
  - $\mathcal{G} = (U, E)$ and its contrast $\mathcal{G}' = (U, E')$
  - matrices $\mathbf{P} = \mathbf{A} + 2\mathbf{D_c}, \mathbf{Q} = \mathbf{A}' + \gamma\mathbf{I}$ related to $\mathcal{G}$ and $\mathcal{G}'$resp.
- **Find**
  - the optimal subset $\boldsymbol{S}^* \subseteq U$,

$$S^* = \underset{S \subseteq V, |S| \geq 1}{\arg\max} g(S; \mathbf{P}, \mathbf{Q}) = \underset{\boldsymbol{x} \in \{0,1\}^n, |\boldsymbol{x}| \geq 1}{\arg\max} \frac{\boldsymbol{x}^T \mathbf{P} \boldsymbol{x}}{\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}}$$



$\boldsymbol{c_i}$: node weight

$\boldsymbol{x}$ is the indicator vector for the node subset

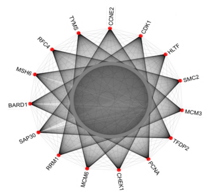
Problems Correspondence

- GENDS and some related instantiation problems

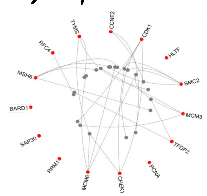
$c_i = -\frac{d_i}{2}$ $\qquad$ $\mathbf{A}' = \mathbf{0},\ \gamma = 1$

	Method	matrix P	matrix Q
1	MinQuotientCut	$\mathbf{A} - \mathbf{D} = -\mathbf{L}$	$\mathbf{I}$
2	Charikar	$\mathbf{A}$	$\mathbf{I}$
3	Fraudar	$\mathbf{A} + 2\mathbf{D}_w$	$\mathbf{I}$
4	SPARSECUTDS [1]	$\mathbf{A} - \frac{2\cdot\alpha}{2\alpha+1}\mathbf{D}$	$\mathbf{I}$
5	TEMPDS	$\mathbf{A}_t$	$\mathbf{A}_{t-1} + 2\mathbf{I} = \tilde{\mathbf{A}}_{t-1}$
6	Risk-averse DS	$\mathbf{A} + \lambda_1\mathbf{I} = \tilde{\mathbf{A}}^{+}$	$\mathbf{A}^{-} + \lambda_2\mathbf{I} = \tilde{\mathbf{A}}^{-}$
	GENDS [2]	$\mathbf{A} + 2\mathbf{D}_c$	$\mathbf{A}' + \gamma\mathbf{I} = \tilde{\mathbf{A}}'$

Graph at time *t*



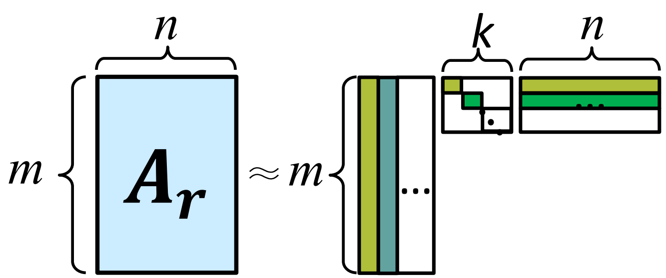
Graph at time *t-1*, $\gamma = 2$



Theoretical Analysis

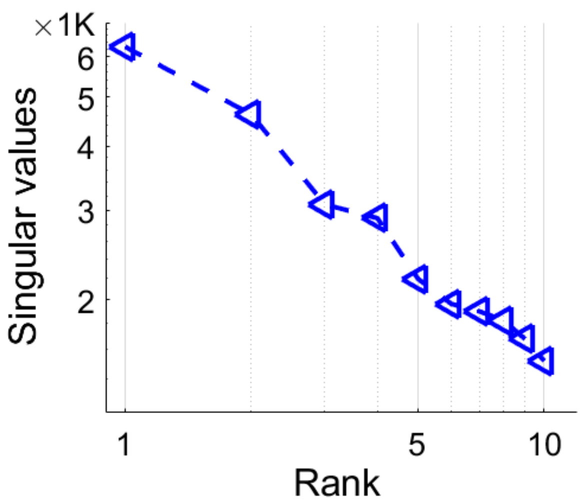
- 1. Graph spectral and Optimality
  - Generalized *Rayleigh ratio* (quadratic optimization)
    - Positive residual graph $\mathcal{G}_r$: $\boldsymbol{A}_r = (\boldsymbol{P} - \boldsymbol{Q})^+$
    - $R(\boldsymbol{A}_r, \boldsymbol{x}) = \frac{\boldsymbol{x}^T \mathrm{A}_r \mathrm{x}}{\boldsymbol{x}^T \boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}$
  - *Rayleigh-Ritz* Theorem (real space)
    - $\boldsymbol{A}_r = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \boldsymbol{\Lambda} = diag(\lambda_1, \cdots \lambda_n)$
    - $\lambda_k = \max_{\boldsymbol{x} \neq \boldsymbol{0},\, \boldsymbol{x} \in \mathcal{S}_{k-1}^{\perp}} R(\mathbf{A}_r, \boldsymbol{x}) = \max_{\|\boldsymbol{x}\|=1,\, \boldsymbol{x} \in \mathcal{S}_{k-1}^{\perp}} \boldsymbol{x}^T \mathbf{A}_r \boldsymbol{x} \Longrightarrow \boldsymbol{x} = \boldsymbol{u}_k$
  - $\boldsymbol{A}_r = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \boldsymbol{\Sigma} = diag(\sigma_1, \ldots, \sigma_k)$

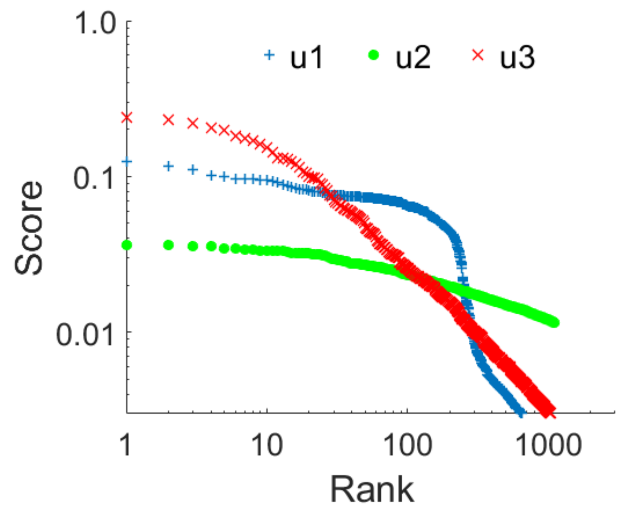
    $\sigma_1 \geq \cdots \geq \sigma_k \geq 0$



Theoretical Analysis (cont.)

- 2. **Properties of real-world large graphs**
  - Sparse connection
  - Skew distributions (e.g., power-law)
    - degree, eigen/singular value, network values, |bipartite core| etc.





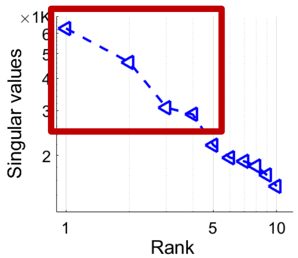
Singular values and score of singular vector elem. distribution of **soc-twitter** network

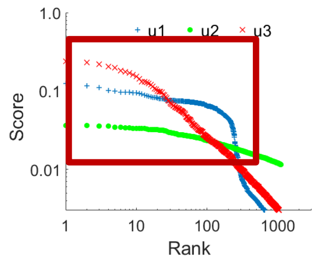
Theoretical Analysis: Optimal Solution

- GENDS **solution** for $\mathcal{G}_r$
  - $$S^* = \arg\max_{\boldsymbol{x}\in\{0,1\}^n, |\boldsymbol{x}|\geq 1} \frac{\boldsymbol{x}^T \mathbf{A}_r \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \arg\max_{|S|\geq 1} \frac{1}{|S|} \sum_{i=1}^{n} \sigma_i \left(\sum_{j\in S} \boldsymbol{u}_{ij}\right)\left(\sum_{j\in S} \boldsymbol{v}_{ij}\right)$$
  - The optimal density of subgraph of S^*:

    $$\boldsymbol{\rho_{opt} \leq \sigma_1}$$

  - **Important parts**:
    - top singular values
    - top-ranked elements in singular vectors

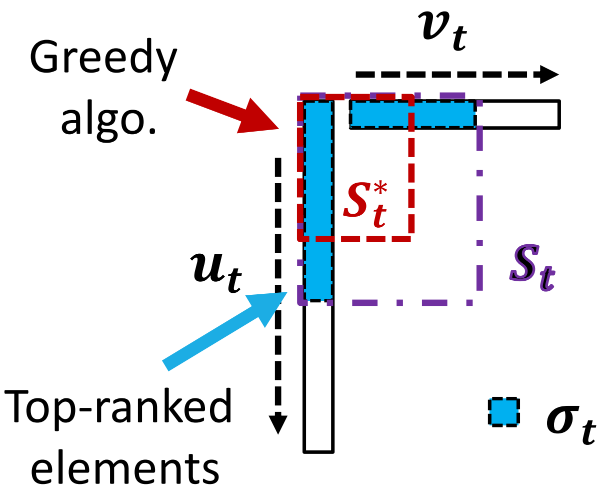


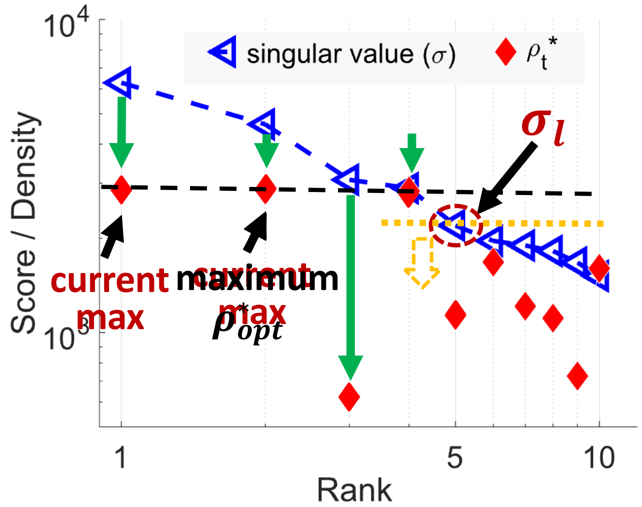


Road Map

- Introduction
- Proposed Method:
  - Unified Formulation: GENDS
  - Algorithm: **SPECGREEDY** <<
- Experiments
- Conclusion

Approximate Optimal Solution





- Approximation solution S_t^*:
  - $\mathcal{g}_t$: subgraph for S_t of the trunc. t-th top singular vectors (u_t, v_t)
  - $\mathcal{g}_t^*$: the densest subgraph of $\mathcal{g}_t$ with Greedy algo.
  - Density of $\mathcal{g}_t^*$: $\boldsymbol{\rho_t^*} \preccurlyeq \sigma_t$
- S_{opt}^* approx. optimal solution for $\mathcal{G}_r$: $\sigma_1 > \cdots > \boldsymbol{\rho_{opt}^*} \succcurlyeq \sigma_l \geq \cdots$

SPECGREEDY: Unified Detection Algorithm

- Graph spectral + Greedy peeling approx.
- Main steps: (k: parameter for SVD approx.)
  - S0: construct positive residual graph $\boldsymbol{\mathcal{G}_r}$
  - S1: $\boldsymbol{A_r}$ SVD (top-k low-rank approx.)
  - *Repeat*:

S2
  - S2-1: Construct candidate $\boldsymbol{S_t}$ for truncated $(\boldsymbol{u}_t, \boldsymbol{v}_t)$
  - S2-2: $\boldsymbol{S_t^*}$: densest subgraph with greedy detection for S_t
  - S2-3: **Update** current optimal density $\boldsymbol{\rho_{opt}^*}$ with S_t^* if needed

  - *Until* $\boldsymbol{\rho_{opt}^*} > \boldsymbol{\sigma_{t+1}}$ or $t < k$
  - *Return* $\boldsymbol{\mathcal{G}_r}(S_{opt}^*)$

Road Map

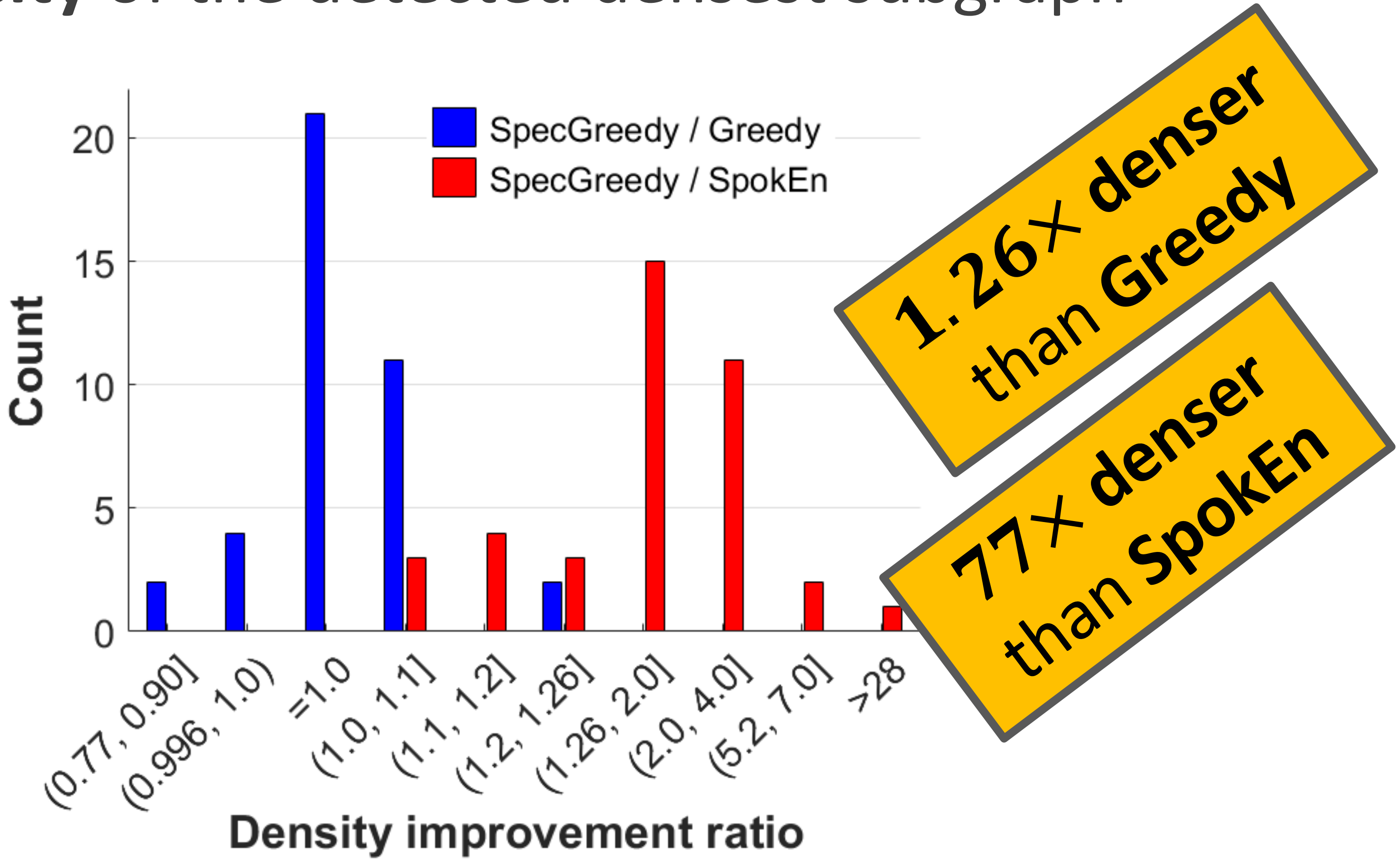
- Introduction
- Proposed Method:
  - Unified Formulation: GenDS
  - Algorithm: SpecGreedy
- Experiments <<
- Conclusion

Experiments on 40 networks

- Baselines:
  - Greedy (Charikar)
  - SpokEn (SVD)
- Parameter: $k = 10$ for SPECGREEDY
- Datasets:
  - **40** real-world networks from **5** public dataset repos.
  - Stanford SNAP, Network Repository, AMiner scholar, ASU' social computing, KONECT Network Collection

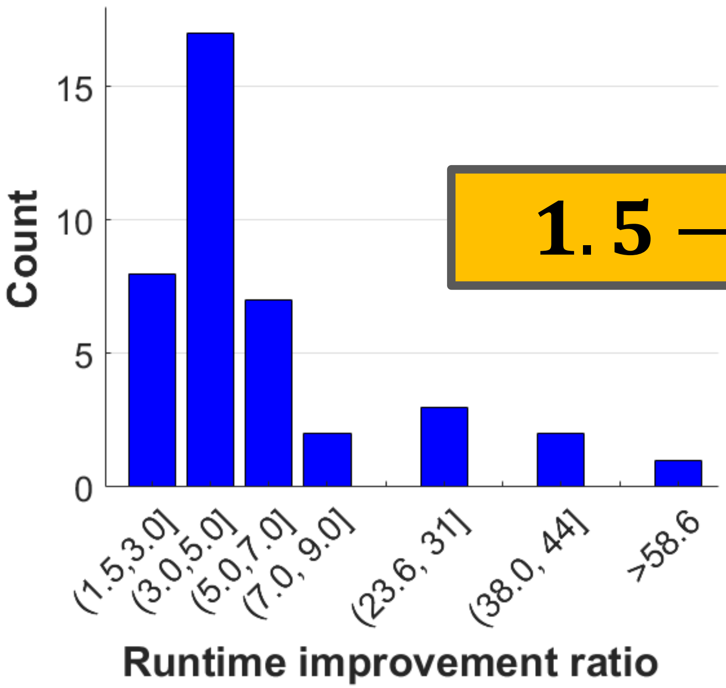
Exp1. Quality-Speed Test

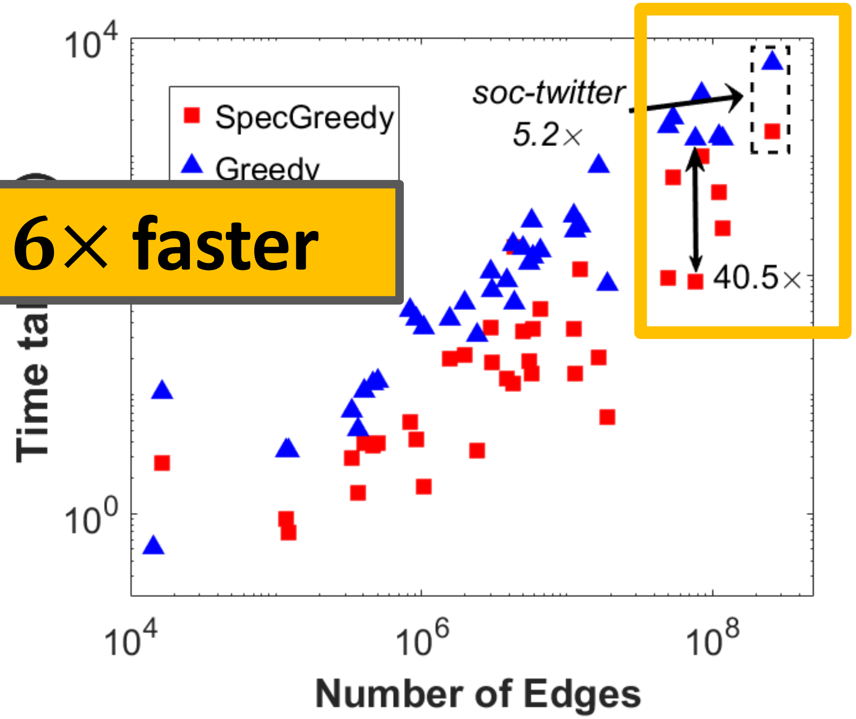
- The **density** of the detected densest subgraph



Exp1. Quality-Speed Test (cont.)

- The **speed-up** for detecting the densest subgraph
  - SpecGreedy vs. Greedy

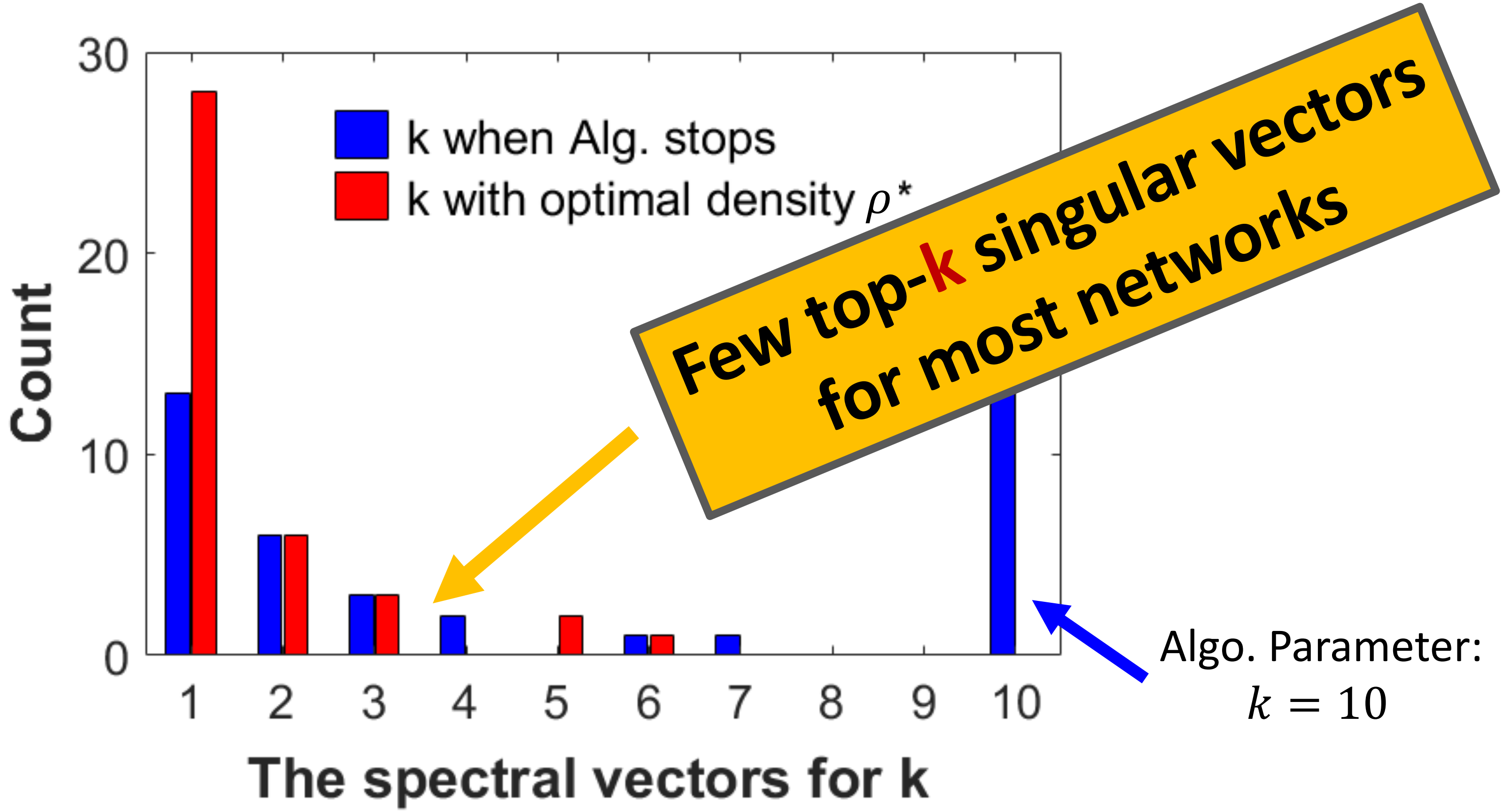




$1.5 - 58.6\times$ faster

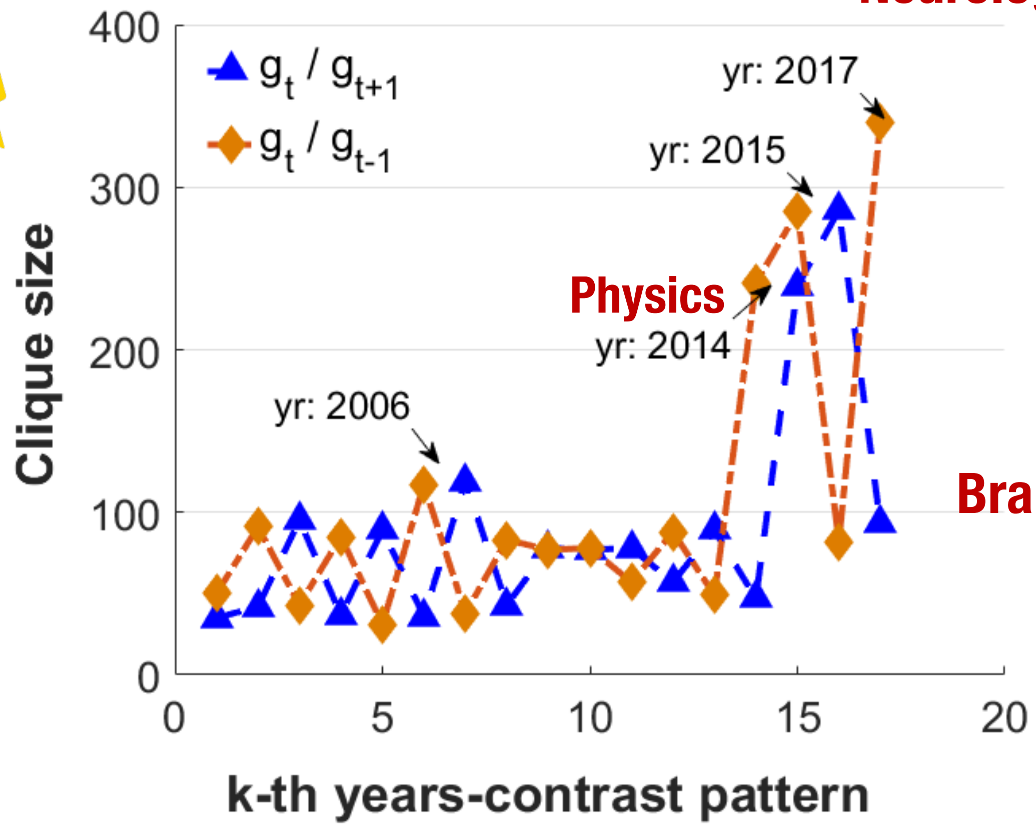
Exp2. Parameter Test

- Optimal graph spectral for the densest subgraph



Exp3. Effectiveness Test

- **Contrast dense** subgraphs in dynamic graph



Neurology & Medicine

Memory, executive, and multidomain subtle cognitive impairment: clinical and biomarker findings

Jon B Toledo [1], Maria Bjerke [1], Kewei Chen [1], Martin Rozycki [1], Clifford R Jack Jr [1], Michael W Weiner [1], Steven E Arnold [1], Eric M Reiman [1], Christos Davatzikos [1], Leslie M Shaw [1], John Q Trojanowski [2], Alzheimer's Disease Neuroimaging Initiative

Collaborators, Affiliations – collapse

Collaborators

Alzheimer's Disease Neuroimaging Initiative: Michael Weiner, Paul Aisen, Michael Weiner, Paul Aisen, Ronald Petersen, Clifford R Jack Jr, William Jagust, John Q Trojanowki, Arthur W Toga, Laurel Beckett, Robert C Green, Andrew J Saykin, John Morris, Enchi Liu, Robert C Green, Tom Montine, Ronald Petersen, Paul Aisen, Anthony Gamst, Ronald G Thomas, Michael Donohue, Sarah Walter, Devon Gessert, Tamie Sather, Laurel Beckett, Danielle Harvey, Anthony Gamst, Michael Donohue, John Kornak, Clifford R Jack Jr, Anders Dale, Matthew Bernstein, Joel Felmlee, Nick Fox, Paul Thompson, Norbert Schuff, Gene Alexander, Charles DeCarli, William Jagust, Dan Bandy, Robert A Koeppe, Norm Foster, Eric M Reiman, Kewei Chen, Chet Mathis, John Morris, Nigel

Brain network and Disease

Brain amyloidosis ascertainment from cognitive, imaging, and peripheral blood protein measures

Liana G Apostolova [1], Kristy S Hwang [2], David Avila [2], David Elashoff [2], Omid Kohannim [2], Edmond Teng [2], Sophie Sokolow [2], Clifford R Jack [2], William J Jagust [2], Leslie Shaw [2], John Q Trojanowski [2], Michael W Weiner [2], Paul M Thompson [2], Alzheimer's Disease Neuroimaging Initiative

Collaborators, Affiliations – collapse

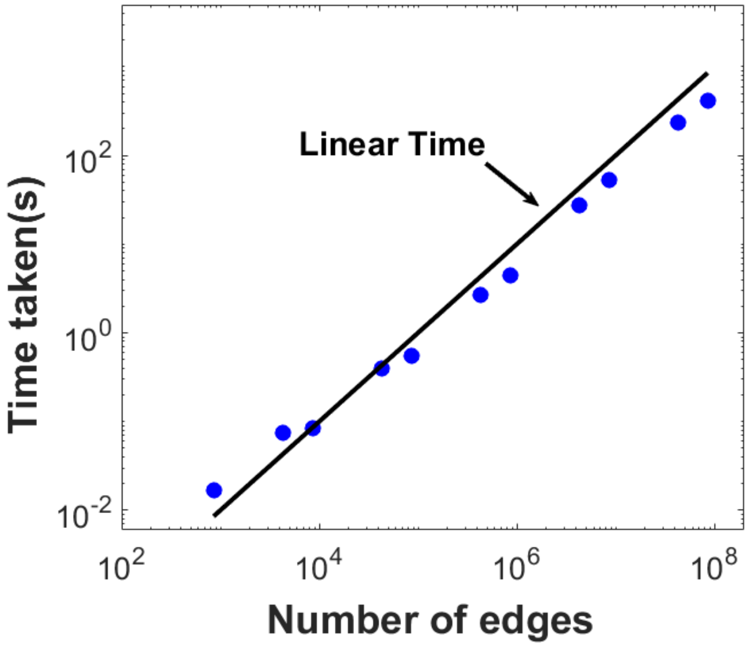
Collaborators

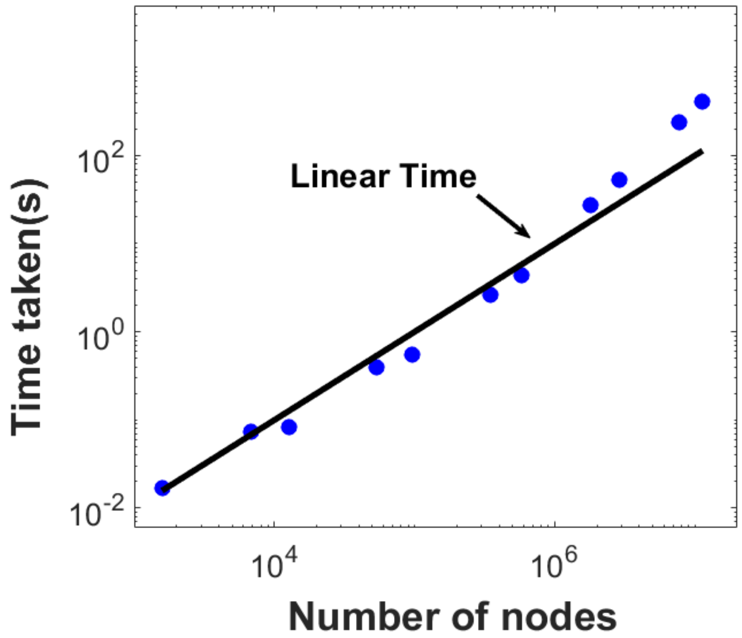
Alzheimer's Disease Neuroimaging Initiative: Michael Weiner, Paul Aisen, Ronald Petersen, Clifford R Jack Jr, Wiliam Jagust, J Q Trojanowki, Arthur W Toga, Laurel Beckett, Robert C Green, Anthony Gamst, Andrew J Sakin, John Morris, William Z Potter, Tom Montine, Michael Donohue, Sarah Walter, Anders Dale, Matthew Bernstein, Joel Felmlee, Nick Fox, Paul Thompson, Norbert Schuff, Gene Alexander, Charles DeCarli, Dan Bandy, Robert A Koeppe, Norm Foster, Eric M Reiman, Kewei Chen, Chet Mathis, Nigel J Cairns, Lisa Taylor-Reinwald, Lee Shaw, Virginia M-Y Lee, Magdalena Korecka, Karen Crawford, Scott Neu, Danielle Harvey, John Kornak, Tatiana M Foroud, Steven Potkin, Li Shen, Zaven Kachaturian, Richard Frank, Peter J

2017 yr, 300+ co-authors

Exp4. Scalability Test

- Time complexity: $O\left(K \cdot |E| + K \cdot \left|E(\tilde{S})\right| \log\left|\tilde{S}\right|\right)^{1}$
- Linear scalable with the size of the graph





[1] $\tilde{S}$: max-size candidate S_t, $\left|\tilde{S}\right| \ll |S|$

Experiments

Properties of **SPECGREEDY:**

- ☑ ***Faster:*** detects the high-quality the densest subgraphs faster than the Greedy algo.
- ☑ ***Effective***: detects the contrast dense subgraph pattern in scholar co-authorship
- ☑ ***Scalable***: runs linearly w.r.t graph size

Road Map

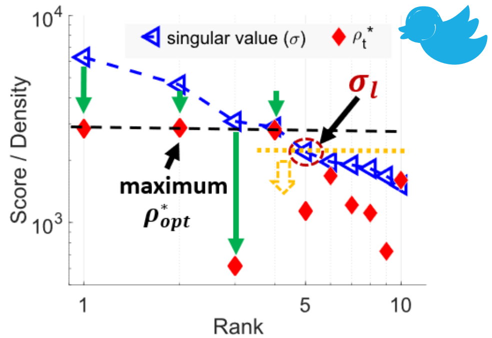
- Introduction
- Proposed Method:
  - Unified Formulation: GENDS
  - Algorithm: SPECGREEDY
- Experiments
- Conclusion <<

Summary

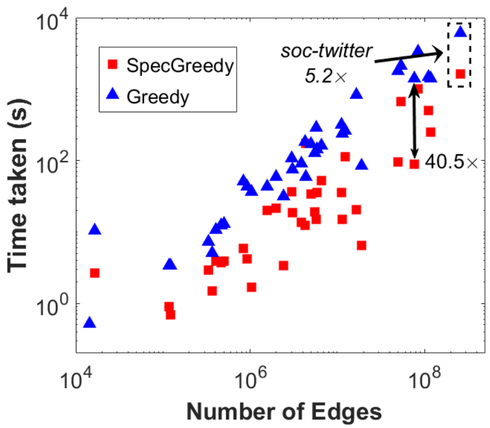
- **SPECGREEDY**: Unified Dense Subgraph Detection

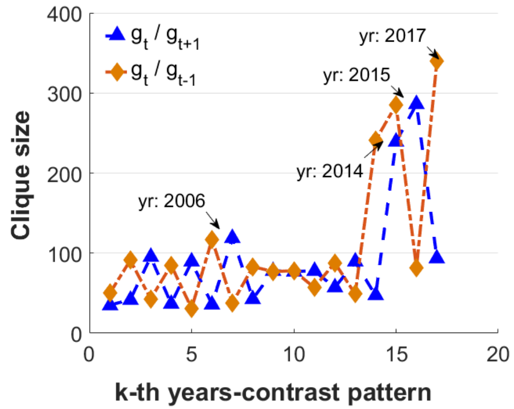
Method	matrix $\mathbf{P}$	matrix $\mathbf{Q}$
MinQuotientCut [9]	$\mathbf{A} - \mathbf{D} = -\mathbf{L}$	$\mathbf{I}$
Charikar [6]	$\mathbf{A}$	$\mathbf{I}$
Fraudar [14]	$\mathbf{A} + 2\mathbf{D}_w$	$\mathbf{I}$
SPARSECUTDS [1] [24]	$\mathbf{A} - \frac{2\cdot\alpha}{2\alpha+1}\mathbf{D}$	$\mathbf{I}$
TEMPDS [33]	$\mathbf{A}_t$	$\mathbf{A}_{t-1} + 2\mathbf{I} = \tilde{\mathbf{A}}_{t-1}$
Risk-averse DS [30]	$\mathbf{A}^{+} + \lambda_1\mathbf{I} = \tilde{\mathbf{A}}^{+}$	$\mathbf{A}^{-} + \lambda_2\mathbf{I} = \tilde{\mathbf{A}}^{-}$
GENDS [2]	$\mathbf{A} + 2\mathbf{D}_c$	$\mathbf{A}' + \gamma\mathbf{I} = \tilde{\mathbf{A}}'$

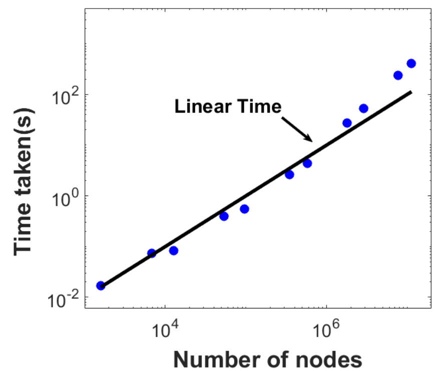
Unified formulation (GenDS)



Graph spectral guaranty







☑ ***Faster*** ☑ ***Effective*** ☑ ***Scalable***

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

M EECS

Thank You!

Visit our paper to learn more

papers

code

Questions:
wenchiehfeng.us@gmail.com

REFERENCE

- Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: bounding graph fraud in the face of camouflage. KDD 2016.
- C. E. Tsourakakis, T. Chen, N. Kakimura, and J. W. Pachocki. Novel dense subgraph discovery primitives: risk aversion and exclusion queries. ECML-PKDD, 2019.
- Fan. R. K. Chung. Spectral graph theory[B]. American mathematical soc., 1996.
- M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In APPROX, 2000.
- Miyauchi, Atsushi, and Naonori Kakimura. "Finding a dense subgraph with sparse cut.“, CIKM. 2018.
- R. Andersen and K. Chellapilla. Finding dense subgraphs with size bounds. In WAW, 2009.
- S. Liu, B. Hooi, and C. Faloutsos. A contrast metric for fraud detection in rich graphs. TKDE, 2018.
- Y. Yang, L. Chu, Y. Zhang, Z. Wang, J. Pei, and E. Chen. Mining density contrast subgraphs. In ICDE, IEEE, 2018.
- Y. Asahiro, K. Iwama, H. Tamaki, and T. Tokuyama. Greedily finding a dense subgraph. Journal of Algorithms, 2000.