Overlapping community detection via constrained PARAFAC: A divide and conquer approach

Fatemeh Sheikholeslami, and Georgios B. Giannakis

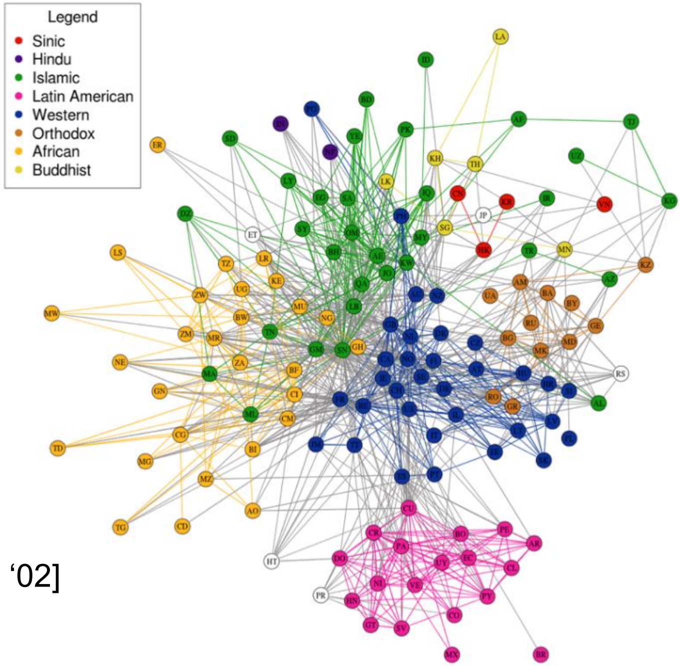
Acknowledgement
NSF grants 1442686, 1500713, 1514056, and NIH grant 1R01GM104975-01

DTC and Dept. of ECE, University of Minnesota

ICDM 2017
NEW ORLEANS

Learning over networks

- Social, biological, and financial networks can be represented by graphs
- Graph $\mathcal{G}$ is given by set of vertices and edges (V, E)
  - $|V| = N$
  - (Un)directed edges $E \subset V \times V$



- Real-world graphs exhibit common properties
  - Power-law degree distribution [Faloutsos et al.]
  - Small-world phenomenon [Barthelemy and Amaral '99]
  - Subgroups with dense connectivity [Girvan and Newman '02]

Interpreted as communities

Natural emergence of communities in the top 1000 country-country Email frequency over 10 million Emails- Washington Post, 2012.

S. Fortunato, Community detection in graphs, *Physics reports*, vol.486, no. 3, pp.75-174, 2010.

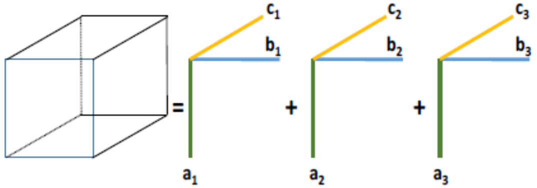
Community detection

- Graph partitioning [Karypis et al.'98, He et al. '15, Whang et al. '15]
- Local greedy algorithms [Derenyi et al. '05, Whang et al. '13]
- Nonnegative matrix fact. [Wang et al. '11, Lesckovec et al. 13, Cao et al. '13, Baingana et al. '16]
- Modularity optimization [Duch et al. '05, Blondel et al. '08]
- Tensor-based graph analysis
  - Dynamic networks [Koutra et al. '12, Araujo et al. '14]
  - Multi-view networks [Papalexakis et al. '13, Araujo et al. '17]
  - Higher-order structures [Huan et al. '15, Benson et al. '15]
- See survey

  S. Fortunato and H. Darko, "Community detection in networks: A user guide." *Physics Reports,* no. 659, pp. 1-44, 2016

Can tensors increase robustness, when only graph adjacency is given?

Tensor decomposition



- Consider data tensor $\underline{X} : I \times J \times N$
- PARAFAC/CPD Tensor model: $\underline{X} = \sum_{f=1}^{F} \boldsymbol{a}_f \circ \boldsymbol{b}_f \circ \boldsymbol{c}_f$

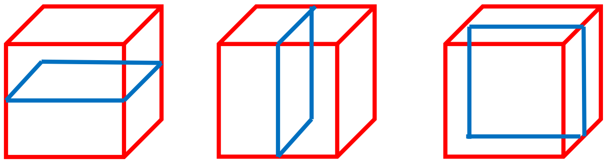
$$\mathbf{A} := [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_F] : I \times F \quad \mathbf{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_F] : J \times F \quad \mathbf{C} := [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_F] : N \times F$$

$$\boxed{\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\underline{\mathbf{X}} - \sum_{f=1}^{F} \boldsymbol{a}_f \circ \boldsymbol{b}_f \circ \boldsymbol{c}_f\|_F^2}$$

Non-convex!

- Can be solved using alternating optimization / Gradient descent (GD/SGD) etc.

Matrix views of $\underline{X}$: $\begin{cases} \mathbf{X}^{(1)} = (\mathbf{C} \odot \mathbf{B})\,\mathbf{A}^\top : KN \times I \\ \mathbf{X}^{(2)} = (\mathbf{C} \odot \mathbf{A})\,\mathbf{B}^\top : IN \times J \\ \mathbf{X}^{(3)} = (\mathbf{B} \odot \mathbf{A})\,\mathbf{C}^\top : IJ \times N \end{cases}$



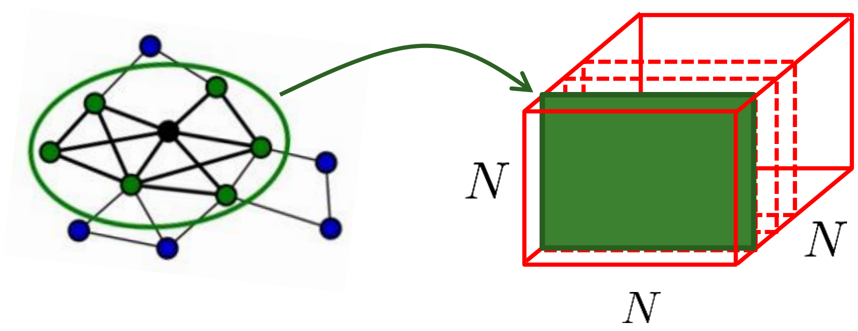
- Off-the-shelf solvers: *N-way toolbox, Tensorlab, AO-ADMM, ParCube, SPLATT*

R. A. Harshman and M. E. Lundy, "PARAFAC: Parallel factor analysis," *Computational Statistics & Data Analysis*, vol. 18, no. 1, pp. 39–72, 1994.

N.D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E.E. Papalexakis, and C. Faloutsos, ''Tensor Decomposition for Signal Processing and Machine Learning'', IEEE Transactions on Signal Processing vol. 65, no. 13, pp. 3551-3582, 2017.

Tensor of egonets

- **Goal:** community detection over graphs using higher order nodal stats.
- Egonets
  - Subgraphs comprising a node and its neighbors



Algorithm Egonet-tensor construction

procedure EGONET-TENSOR CONSTRUCTION($\mathcal{V}, \mathbf{W}$)
  for $n \in \mathcal{V}$ **do**
    $\mathcal{N}(n) := \{v \in \mathcal{V} | w_{nv} \neq 0\}$
    $\mathbf{W}^n \leftarrow \text{subgraph}\Big(\{n\} \cup \mathcal{N}(n), \mathbf{W}\Big)$
    $\underline{\mathbf{W}}_{:,:,n} = \mathbf{W}^{(n)}$
  end for
end procedure
return $\underline{\mathbf{W}}$

- Egonet-tensor provides an **enhanced** 3-D network representation

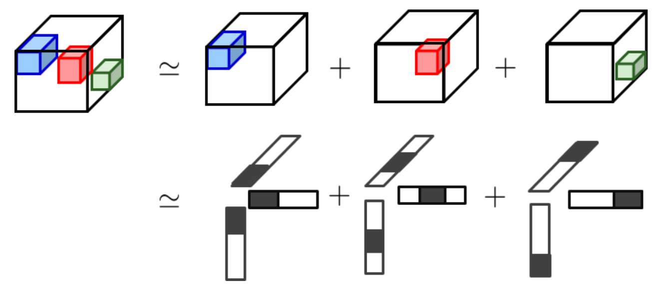
L. Akoglu, M. McGlohon, and C. Faloutsos. “Oddball: Spotting anomalies in weighted graphs.” In *Advances in Knowledge Discovery and Data Mining*, pp. 410-421, 2010.

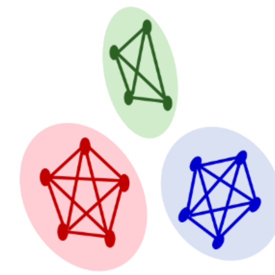
Community patterns in egonet-tensors

- Network of densely-connected and nonoverlapping communities

Q: How does the corresponding ***egonet-tensor*** look like?

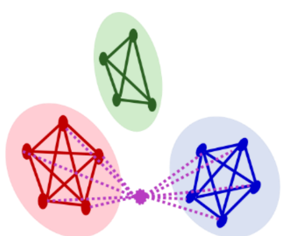
$\underline{\mathbf{W}}$:



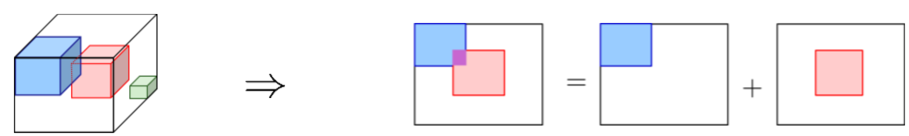


CPD model over egonet- tensor

Q: How does the egonet-tensor look like with ***overlapping communities***?



$\underline{\mathbf{W}}$:



Q: How does the egonet-tensor look like in **real-world networks**?

✓ **Dense blocks with structured redundancy** $\Rightarrow$ **increased robustness!**

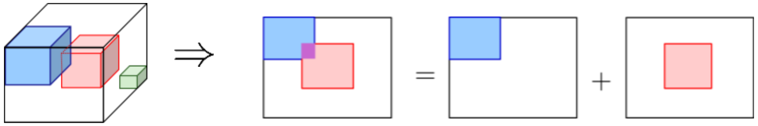
PARAFAC of egonet-based tensor

- Nonnegativity constraint

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\underline{\mathbf{W}} - \sum_{f=1}^{K} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f\|_F^2$$
$$\text{s.t.} \quad \mathbf{A} \geq 0,\ \mathbf{B} \geq 0,\ \mathbf{C} \geq 0$$

- Simplex constraint over community association

$$\underline{\mathbf{W}}(:,:,n) = \sum_{f=1}^{K} c_{nf}\underbrace{(\mathbf{a}_f \circ \mathbf{b}_f)}_{\text{Adjacency pattern for com. } f}$$



Node *n* assoc. to com. *f*

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\underline{\mathbf{W}} - \sum_{f=1}^{K} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f\|_F^2 + \lambda(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$$
$$\text{s.t.} \quad \mathbf{A} \geq 0,\ \mathbf{B} \geq 0,\ \mathbf{C} \geq 0, \sum_{k=1}^{K} c_{nk} = 1\ \forall n$$

Solver: *ALS with intermediate ADMM iterations*

- Alternating minimization
  - Convex subproblems
  - Exact solution using ADMM iterations

- SPLATT: sparse tensor decomposition toolbox [Smith et al. 2016]

Algorithm Constrained tensor decomposition via ALS

Input $\underline{\mathbf{W}}, K, I_{\max}, \lambda$
Initialize $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{N\times K}$ at random and set $k=0$
Form Matrix reshapes $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ of the tensor as
$\mathbf{W}_1 = [\text{vec}(\text{squeeze}(\underline{\mathbf{W}}_{1,:,:})), \cdots, \text{vec}(\text{squeeze}(\underline{\mathbf{W}}_{N,:,:}))]$
$\mathbf{W}_2 = [\text{vec}(\text{squeeze}(\underline{\mathbf{W}}_{:,1,:})), \cdots, \text{vec}(\text{squeeze}(\underline{\mathbf{W}}_{:,N,:}))]$
$\mathbf{W}_3 = [\text{vec}(\text{squeeze}(\underline{\mathbf{W}}_{:,:,1})), \cdots, \text{vec}(\text{squeeze}(\underline{\mathbf{W}}_{:,:,N}))]$
$\mathbf{H}_A^{(k)} = (\mathbf{C}^{(k-1)} \odot \mathbf{B}^{(k-1)})$
$\mathbf{A}^{(k)} \leftarrow$ algorithm 2 with input $\{\mathbf{H}_A^{(k)}, \mathbf{W}_1, \mathbf{A}^{(k-1)}\}$
$\mathbf{H}_B^{(k)} = (\mathbf{C}^{(k-1)} \odot \mathbf{A}^{(k)})$
$\mathbf{B}^{(k)} \leftarrow$ algorithm 2 with input $\{\mathbf{H}_B^{(k)}, \mathbf{W}_2, \mathbf{B}^{(k-1)}\}$
$\mathbf{H}_C^{(k)} = (\mathbf{B}^{(k)} \odot \mathbf{A}^{(k)})$
$\mathbf{C}^{(k)} \leftarrow$ algorithm 3 with input $\{\mathbf{H}_C^{(k)}, \mathbf{W}_3, \mathbf{C}^{(k-1)}\}$
$k \leftarrow k+1$

- Subproblem for factors $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{Z}^* = \underset{\mathbf{Z}\geq \mathbf{0}}{\arg\min} \text{Tr}\left\{\mathbf{Z}(\mathbf{H}^\top\mathbf{H} + \lambda \mathbf{I}_{K\times K})\mathbf{Z}^\top - 2\mathbf{W}^\top\mathbf{H}\mathbf{Z}^\top\right\}$$

Algorithm ADMM solver for mode-1 and 2 subproblems

Input $\underline{\mathbf{H}}, \mathbf{W}, \mathbf{Z}_{\text{init}}$
Set $\rho = \frac{\|\mathbf{Z}_{\text{init}}\|_F^2}{K}, \mathbf{Z}^{(0)} = \mathbf{Z}_{\text{init}}, \bar{\mathbf{Z}}^{(0)} = \mathbf{0}_{N\times K}, \mathbf{Y}^{(0)} = \mathbf{0}_{N\times K}, r = 0$
while $r < I_{\text{max,ADMM}}$ **do**
$\mathbf{Z}^{(r)} = (\mathbf{H}^\top\mathbf{H} + (\rho+\lambda)\mathbf{I})^{-1}(\mathbf{W}^\top\mathbf{H} + \frac{\rho}{2}(\bar{\mathbf{Z}}^{(r-1)} - \mathbf{Y}^{(r-1)}))$
$\bar{\mathbf{Z}}^{(r)} = \mathcal{P}_+(\mathbf{Z}^{(r)})$
$\mathbf{Y}^{(r)} = \mathbf{Y}^{(r-1)} - \rho(\mathbf{Z}^{(r)} - \bar{\mathbf{Z}}^{(r)})$
$r = r+1$
end while
Retrun $\mathbf{Z}^{(r)}$

- Subproblem for factor $\mathbf{C}$

$$\mathbf{Z}^* = \underset{\mathbf{Z}\geq\mathbf{0}\, \|\tilde{\mathbf{z}}_n\|_1 = 1\ \forall n=1,\cdots,N}{\arg\min} \text{Tr}\left\{\mathbf{Z}\mathbf{H}^\top\mathbf{H}\mathbf{Z}^\top - 2\mathbf{W}^\top\mathbf{H}\mathbf{Z}^\top\right\}$$

Algorithm ADMM solver for mode-3 subproblems

Input $\underline{\mathbf{H}}, \mathbf{W}, \mathbf{Z}_{\text{init}}$
Set $\rho = \frac{\|\mathbf{Z}_{\text{init}}\|_F^2}{K}, \mathbf{Z}^{(0)} = \mathbf{Z}_{\text{init}}, \bar{\mathbf{Z}}^{(0)} = \mathbf{0}_{N\times K}, \mathbf{Y}^{(0)} = \mathbf{0}_{N\times K}, r = 0$
while $r < I_{\text{max,ADMM}}$ **do**
$\mathbf{Z}^{(r)} = (\mathbf{H}^\top\mathbf{H} + \rho\mathbf{I}_{N\times N})^{-1} \times \left(\mathbf{W}^\top\mathbf{H} + \frac{\rho}{2}(\bar{\mathbf{Z}}^{(r-1)} + \mathbf{Y}^{(r-1)})\right)$
$\bar{\mathbf{Z}}^{(r)} = \mathcal{P}_{\text{simp}}(\mathbf{Y}^{(r-1)} + \mathbf{Z}^{(r)})$
$\mathbf{Y}^{(r)} = \mathbf{Y}^{(r-1)} - \rho(\mathbf{Z}^{(r)} - \bar{\mathbf{Z}}^{(r)})$
$r = r+1$
end while
Retrun $\bar{\mathbf{Z}}^{(r)}$

Synthetic tests

- LFR benchmark networks with ground-truth communities $\mathcal{S}^* = \{\mathcal{C}_1^*, \ldots, \mathcal{C}_{|\mathcal{S}|}^*\}$
  - Total number of nodes N
  - Community mixing coefficient μ
  - Number of overlapping nodes o_n and communities o_m

- **Performance metric** for detected communities $\hat{\mathcal{S}} = \{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_{|\hat{\mathcal{S}}|}\}$
  - Normalized Mutual Information (NMI)

$$\mathrm{NMI}(\mathcal{S}^*, \hat{\mathcal{S}}) := \frac{2\mathrm{I}(\mathcal{S}^*, \hat{\mathcal{S}})}{\mathrm{H}(\mathcal{S}^*) + \mathrm{H}(\hat{\mathcal{S}})}$$

$$\mathrm{H}(\hat{\mathcal{S}}) := -\sum_{i=1}^{|\hat{\mathcal{S}}|} p(\hat{\mathcal{C}}_i) \log p(\hat{\mathcal{C}}_i) = -\sum_{i=1}^{|\hat{\mathcal{S}}|} \frac{|\hat{\mathcal{C}}_i|}{N} \log \frac{|\hat{\mathcal{C}}_i|}{N} \qquad \mathrm{I}(\mathcal{S}^*, \hat{\mathcal{S}}) := \sum_{i=1}^{|\mathcal{S}^*|} \sum_{j=1}^{|\hat{\mathcal{S}}|} \frac{|\mathcal{C}_i^* \cap \hat{\mathcal{C}}_j|}{N} \log \frac{N|\mathcal{C}_i^* \cap \hat{\mathcal{C}}_j|}{|\mathcal{C}_i^*||\hat{\mathcal{C}}_j|}.$$

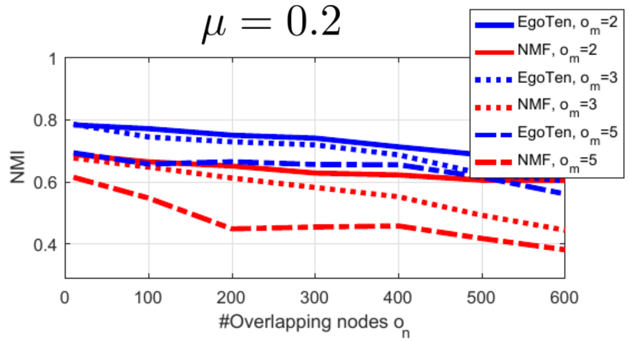
  - F1 score $\bar{F1} := \frac{1}{2|\mathcal{S}^*|} \sum_{i=1}^{|\mathcal{S}^*|} F1(\mathcal{C}_i^*, \hat{\mathcal{C}}_{I(i)}) + \frac{1}{2|\hat{\mathcal{S}}|} \sum_{i=1}^{|\hat{\mathcal{S}}|} F1(\mathcal{C}_{I'(i)}^*, \hat{\mathcal{C}}_i)$ where $F1(\mathcal{C}_i, \mathcal{C}_j) := \frac{2\,|\mathcal{C}_i \cap \mathcal{C}_j|}{|\mathcal{C}_i| + |\mathcal{C}_j|}$.
  - Conductance $\phi(\hat{\mathcal{C}}_k) := \frac{\sum_{i \in \hat{\mathcal{C}}_k, j \notin \hat{\mathcal{C}}_k} \mathbf{W}_{ij}}{\min\{\mathrm{vol}(\hat{\mathcal{C}}_k), \mathrm{vol}(\mathcal{V} \setminus \hat{\mathcal{C}}_k)\}}$ where $\mathrm{vol}(\hat{\mathcal{C}}_k) := \sum_{i \in \hat{\mathcal{C}}_k, \forall j} \mathbf{W}_{ij}$

A. Lancichinetti, S. Fortunato, and F. Radicchi, Benchmark graphs for testing community detection algorithms. Physical Review E, vol. 78, pp. 123-456, 2008.

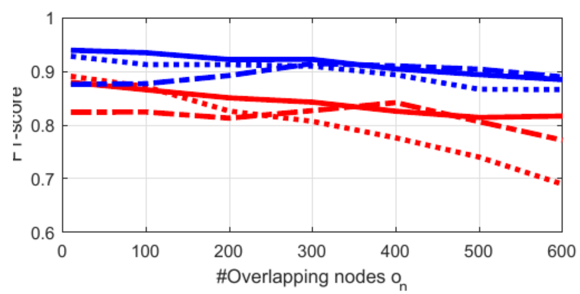
EgoTen vs. NMF

- Benchmark constrained NMF $\min_{\mathbf{U},\mathbf{V}} \|\mathbf{W} - \mathbf{U}\mathbf{V}^\top\|_F^2 \quad \text{s.t. } \|\mathbf{u}_n\|_1 = 1 \, \forall n, \; \mathbf{U} \geq 0 \,, \mathbf{V} \geq 0$
- Synthetic LFR networks with $N = 1000$ and $\bar{d} = 100$

$\mu = 0.2$

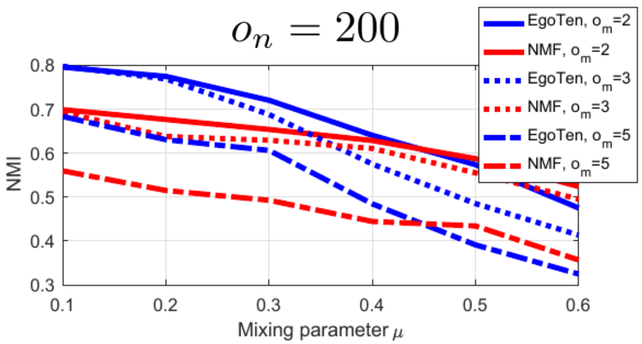


(a)

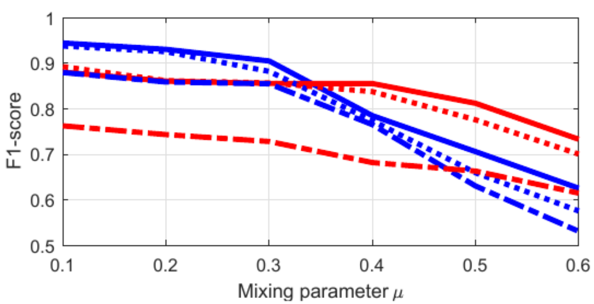


(b)

$o_n = 200$



(a)



(b)

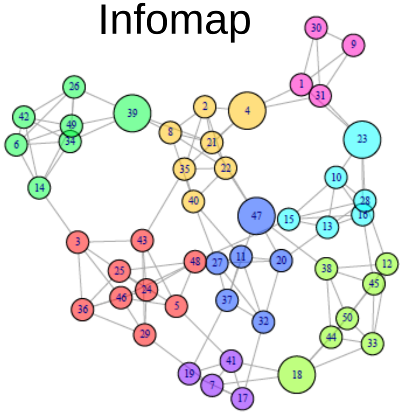
✓ Egonet Tensor representation provides structured redundancy
✓ Increased robustness against overlapping nodes as well as community mixing

F. Sheikholeslami and G. B. Giannakis, "Identification of Overlapping Communities via Constrained Egonet Tensor Decomposition," *IEEE Transactions on Signal Processing*, submitted July 2017.

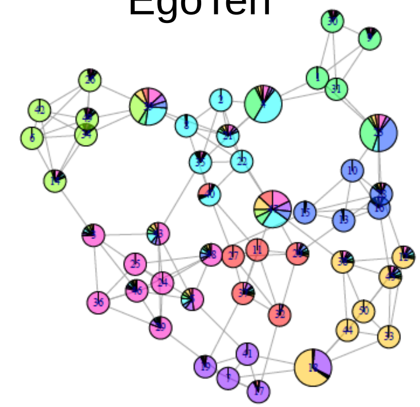
Soft community association

❖ LFR Network

Infomap

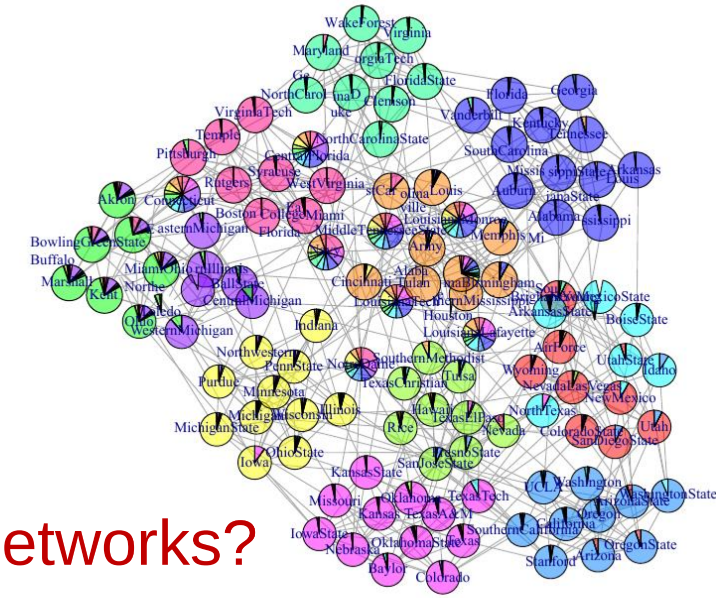


EgoTen



- $N = 50$ nodes
- $o_n = 5$ overlapping nodes
- $\mu = 0.2$ mixing coefficient

❖ American College Football Network



- *N*=119 colleges as nodes
- Nodes connected if the teams compete
- Conferences as communities (*K*=12)

Application on extremely large networks?

Overlapping comm. ID over large-scale networks

❑ State-of-the-art

- Nise [Whang et al.'16],
  Demon [Coscia et al.12],
  BigClam [Yang et al.'13], and more.

❑ Challenges

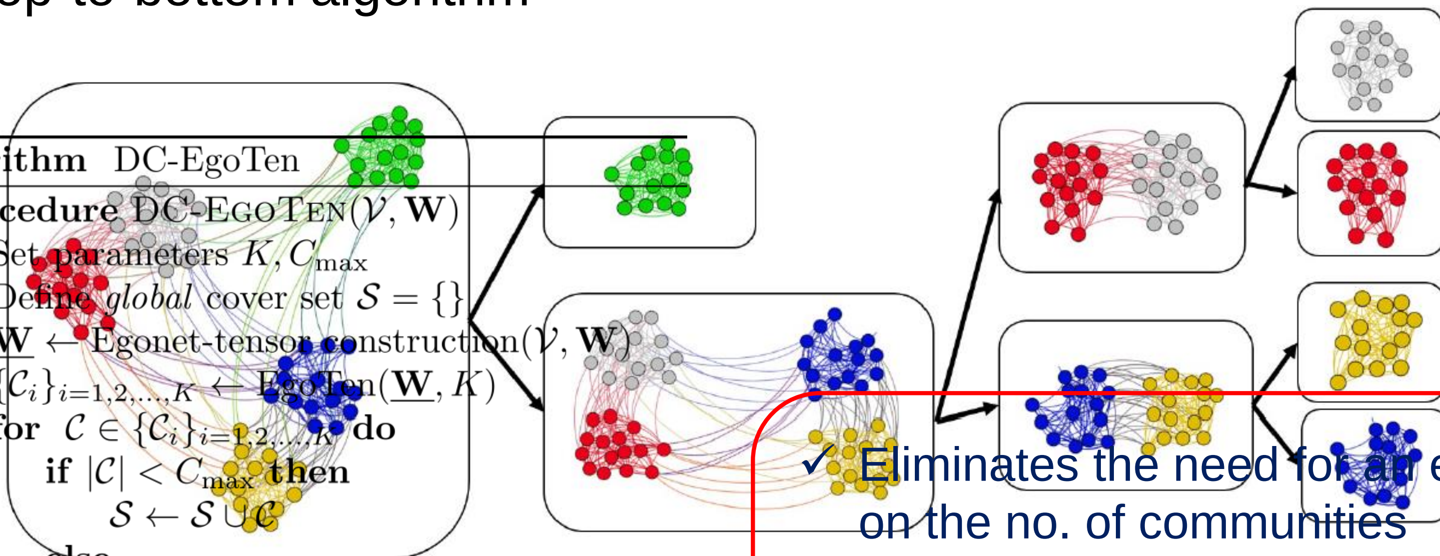
- Scalability
- Quality of detected communities
- Resolution limit

❑ EgoTen considerations

- The need for upperbound on the number of community K
- Tensor decomposition scales well with N, but not so well with K

DC-EgoTen: A divide-and-conquer approach

❖ A top-to-bottom algorithm



Algorithm DC-EgoTen

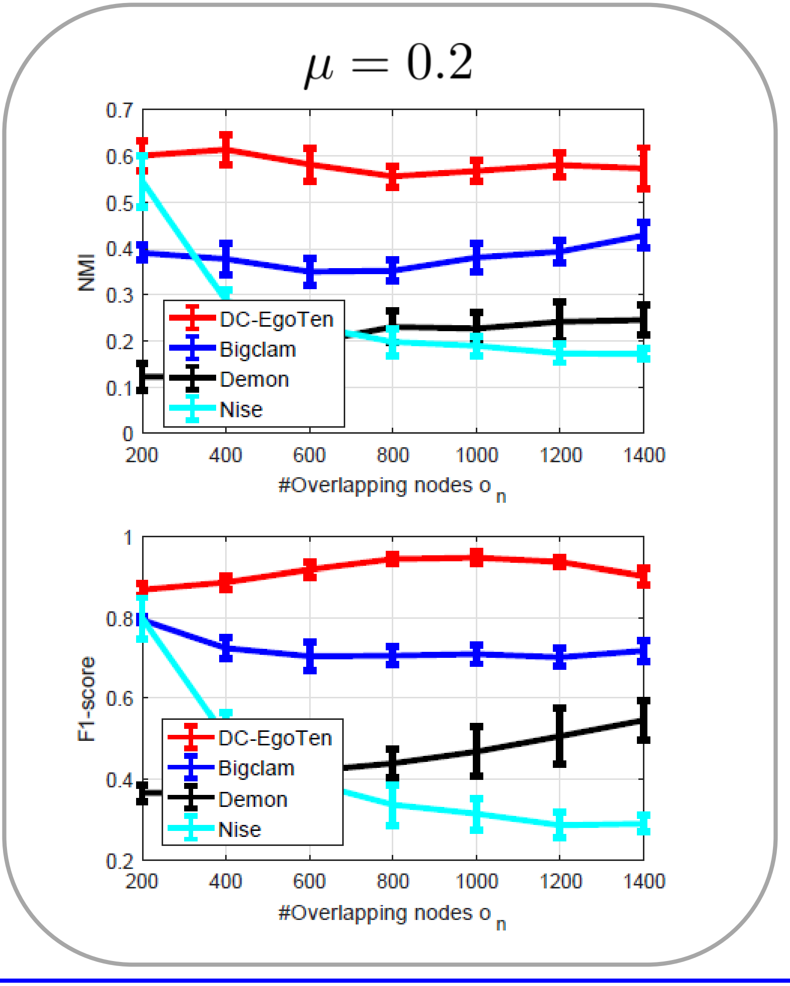
procedure DC-EgoTen($\mathcal{V}, \mathbf{W}$)
  Set parameters $K, C_{\max}$
  Define *global* cover set $\mathcal{S} = \{\}$
  $\underline{\mathbf{W}} \leftarrow$ Egonet-tensor construction($\mathcal{V}, \mathbf{W}$)
  $\{\mathcal{C}_i\}_{i=1,2,\ldots,K} \leftarrow$ EgoTen($\underline{\mathbf{W}}, K$)
  for $\mathcal{C} \in \{\mathcal{C}_i\}_{i=1,2,\ldots,K}$ **do**
    if $|\mathcal{C}| < C_{\max}$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{C}$
    else
      # Extract the subgraph of nodes in $\mathcal{C}$
      $\mathbf{W}_{\text{sub}} \leftarrow$ subgraph($\mathcal{C}, \mathbf{W}$)
      DC-EgoTen ($\mathcal{C}, \mathbf{W}_{\text{sub}}$)
    end if
  end for
end procedure
return $\mathcal{S}$

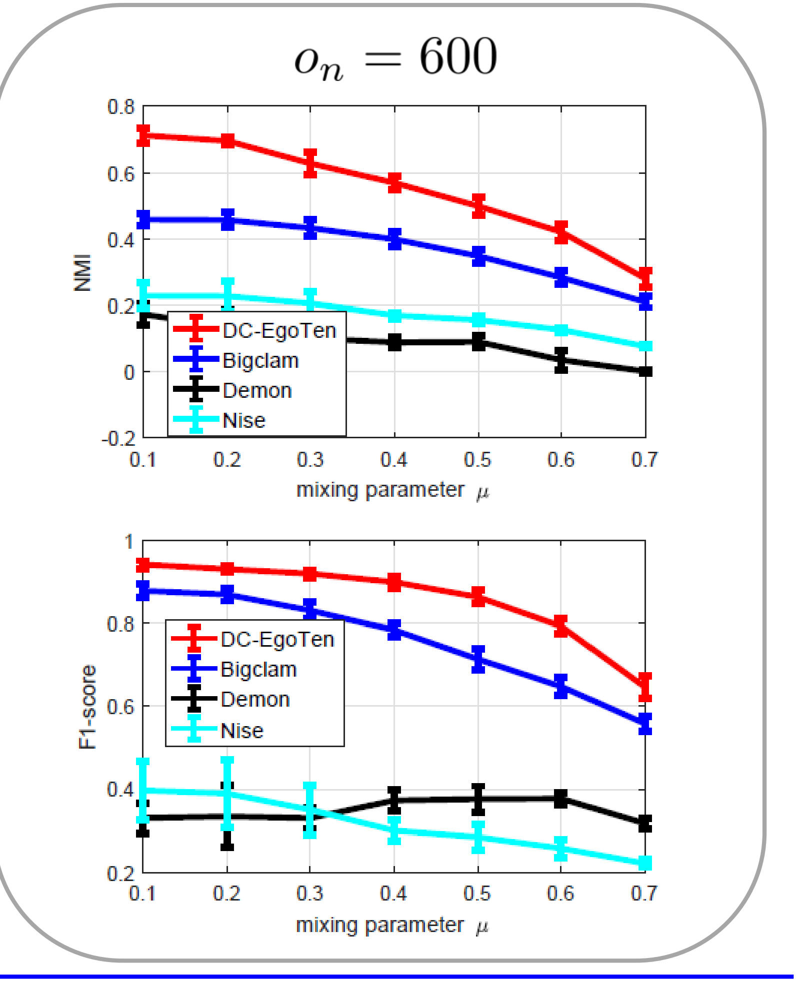
✓ Eliminates the need for an estimation on the no. of communities

✓ Consecutively refines the detected communities up to target resolution

Simulation tests

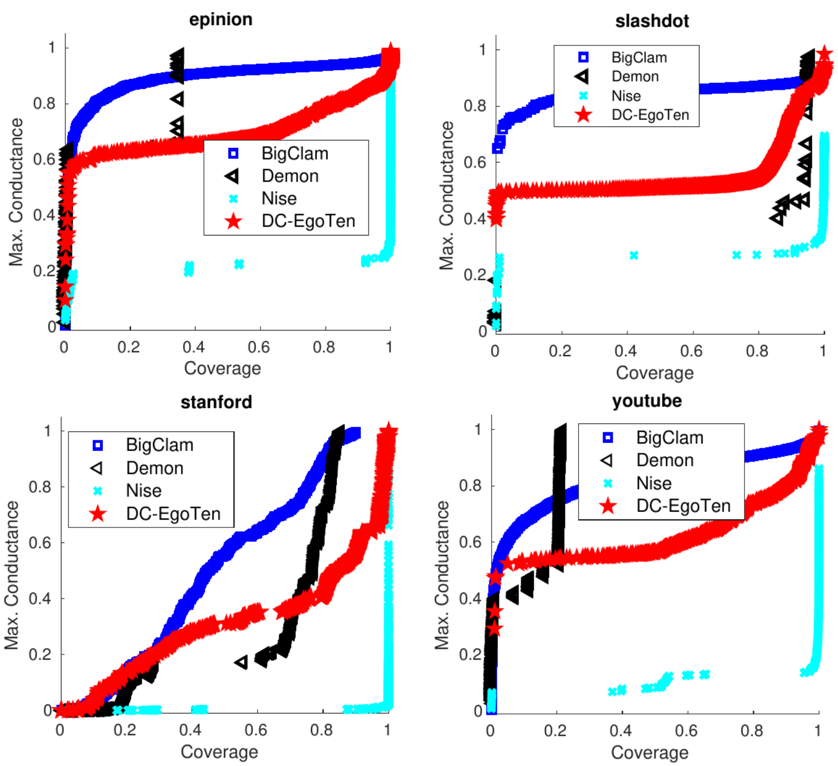
❖ Synthetic LFR networks with $N = 2000$, $\bar{d} = 100$ and $o_m = 3$

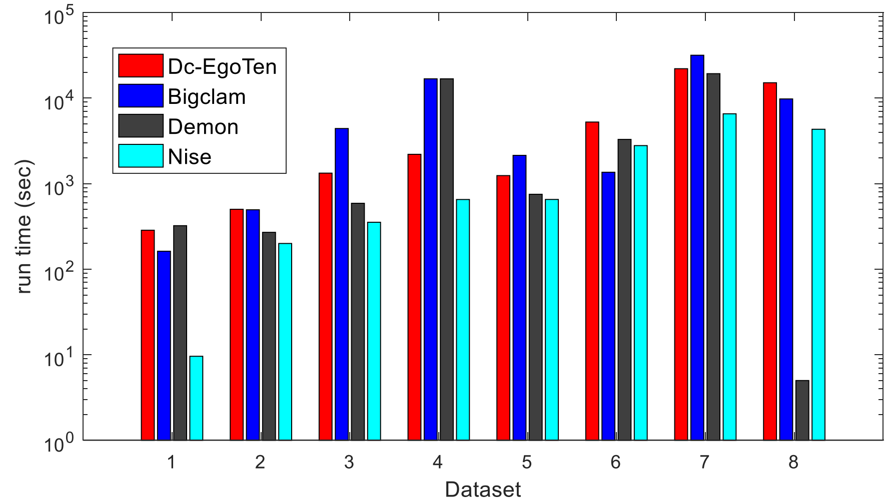




Real-world networks: conductance vs. coverage

Dataset	No. of vertices N	No. of edges $\lvert\mathcal{E}\rvert$	Edge type
Facebook	4,039	88,234	Undirected
Enron	36,692	183,831	Undirected
Epinion	75,879	508,837	Directed
Slashdot	82,168	948,464	Directed
Email	265,214	420,045	Directed
Stanford	281,903	2,312,497	Directed
Notredame	325,729	1,497,134	Directed
Youtube	1,134,890	2,987,624	Undirected





	Dataset		DC-EgoTen	Bigclam	Demon	Nise
1	Facebook	Coverage	100%	95%	99%	89%
		No. of comm.	523	500	8	16
2	Enron	Coverage	100%	90%	65%	100%
		No. of comm.	553	500	343	520
	Slashdot	Coverage	100%	100%	95%	100%
		No. of comm.	1163	500	51	485
4	Epinion	Coverage	100%	100%	35%	100%
		No. of comm.	1274	2000	136	2041
5	Email	Coverage	100 %	83%	11%	100%
		No. of comm.	965	2000	24	2404
6	Notredame	Coverage	100%	100%	39%	100%
		No. of comm.	1169	2000	1497	1454
7	Stanford	Coverage	100%	90%	85%	100%
		No. of comm.	807	2000	2596	1411
8	Youtube	Coverage	100%	100%	22%	100%
		No. of comm.	813	5000	3835	5162

Summary

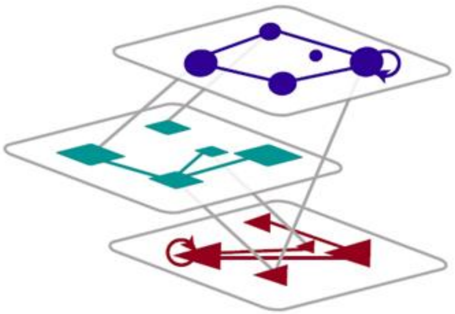
- ❖ Community detection over large networks
- ❖ Egonet-based multi-dimensional network representation
- ❖ Constrained PARAFAC decomposition
- ❖ Top to bottom identification of communities
- ❖ Numerical tests over synthetic and real-world networks
- ❖ Codes available at

https://github.com/FatemehSheikholeslami/EgoTen

Thank you!

Future directions

❖ Multi-layer networks



❖ Unitization of extra-nodal features $\{\mathbf{f}_n\}_{n=1}^{N}$

$$\underset{\mathbf{A},\mathbf{B},\mathbf{C},\{\bar{\mathbf{f}}_k\}_{k=1}^{K}}{\arg\min} \|\underline{\mathbf{W}} - \textstyle\sum_{k=1}^{K} \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k\|_F^2 + \lambda(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2) + \gamma \textstyle\sum_{n=1}^{N} \|\mathbf{f}_n - \textstyle\sum_{k=1}^{K} c_{nk}\bar{\mathbf{f}}_k\|_2^2$$

$$\text{s.t.} \qquad \mathbf{A} \geq \mathbf{0}, \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \qquad \textstyle\sum_{k=1}^{K} c_{nk} = 1 \qquad \forall n = 1, 2, ..., N$$

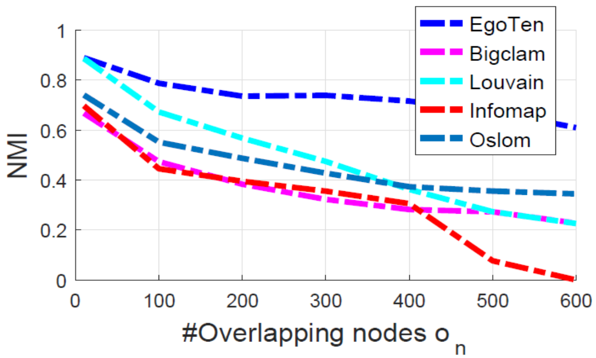
❖ Intruder and anomaly detection

$$\underset{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{O}}{\arg\min} \|\underline{\mathbf{W}} - \textstyle\sum_{k=1}^{K} \mathbf{a}_k \circ \mathbf{b}_k \circ (\mathbf{c}_k + \mathbf{o}_k)\|_F^2 + \nu(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) + \gamma\|\mathbf{O}\|_0$$

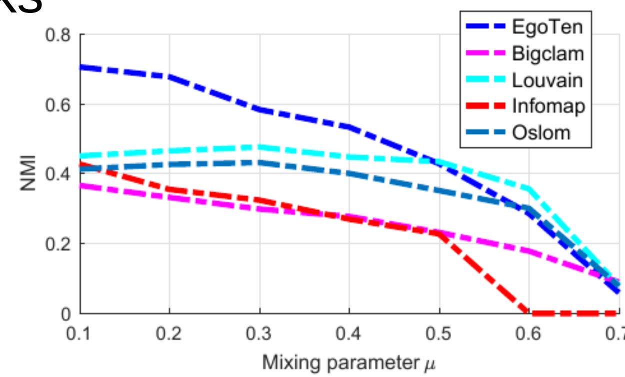
$$\text{s.t.} \qquad \mathbf{A} \geq \mathbf{0}, \qquad \mathbf{B} \geq \mathbf{0}, \qquad \mathbf{C} \geq \mathbf{0}, \qquad \mathbf{O} \geq \mathbf{0}$$

EgoTen vs. state-of-the-art

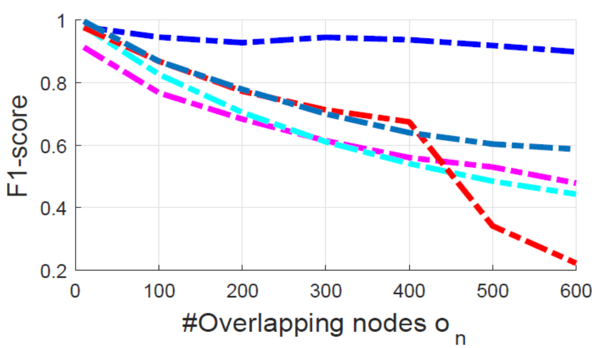
❖ Experiments carried over LFR networks



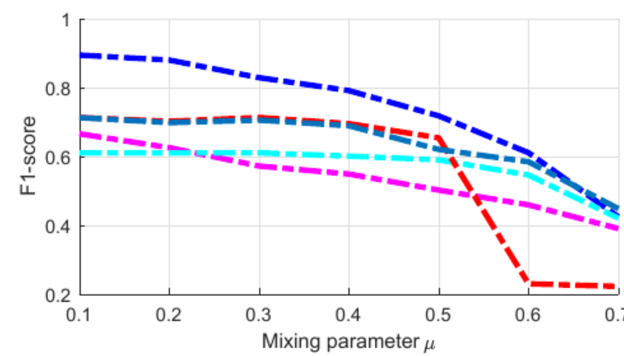
(a)



(a)



(b)

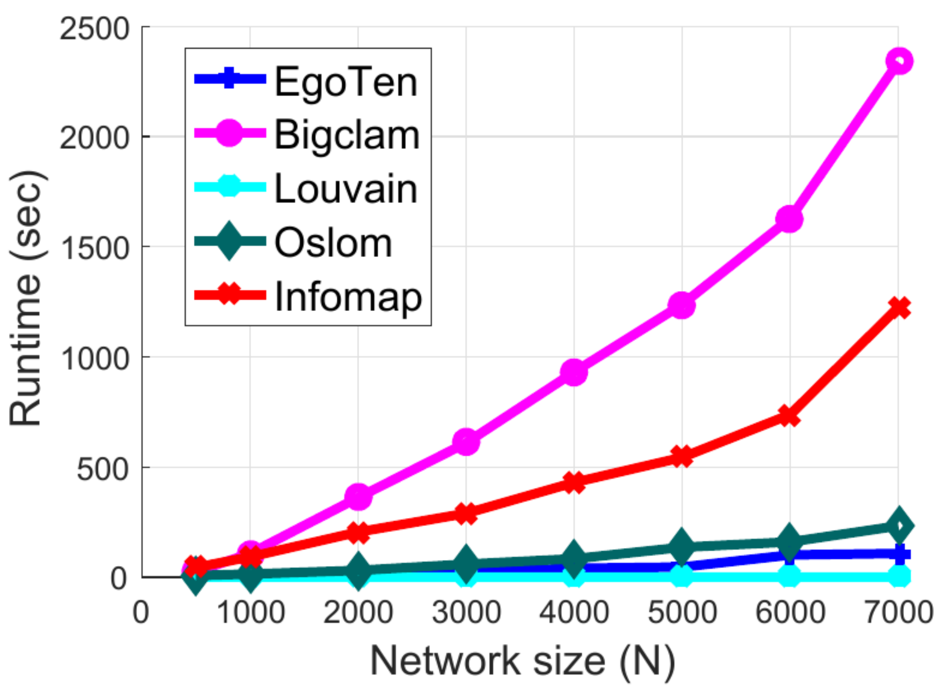


(b)

✓ Egonet Tensor representation provides structured redundancy

✓ Constrained CPD exploits structure for improved robustness

Scalability



Thanks to <u>sparsity</u> and <u>parallelization</u>, EgoTen complexity grows gracefully!