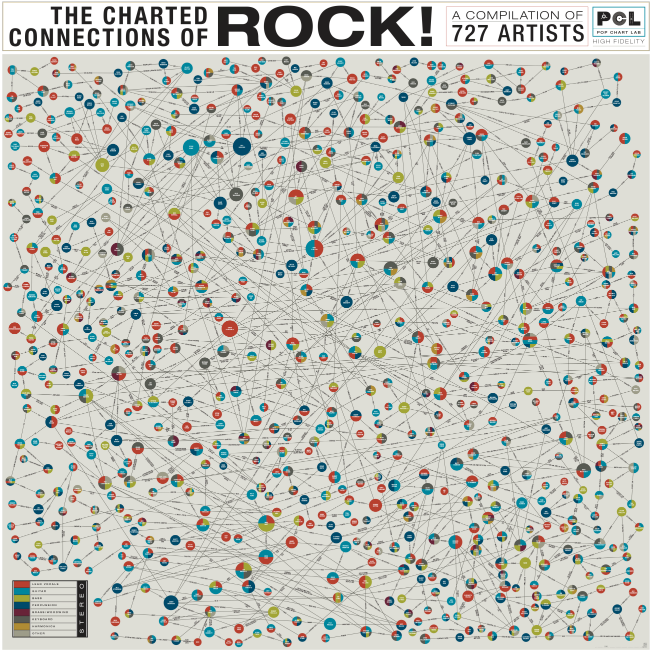
Parsimonious formulations for low-diameter clusters

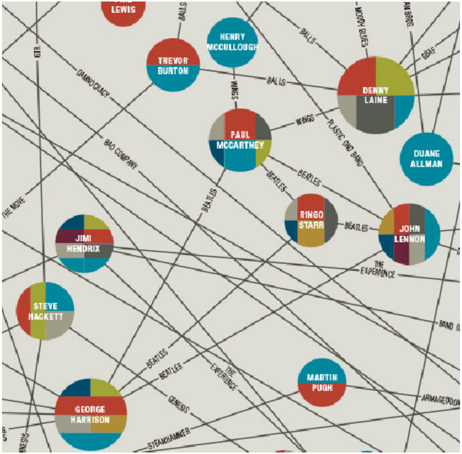
@AustinLBuchanan and @HASalemi

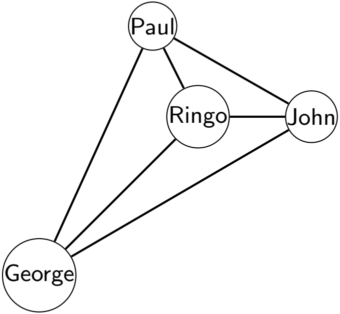
Oklahoma State University
Industrial Engineering & Management

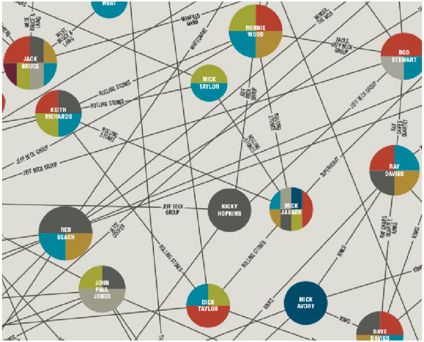
#euro2018valencia

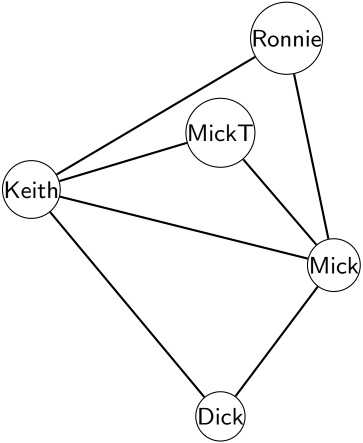
The rock musician graph $G = (V, E)$.




LEWIS
TREVOR BURTON
HENRY MCCULLOUGH
DENNY LAINE
PAUL MCCARTNEY
DUANE ALLMAN
RINGO STARR
JOHN LENNON
JIMI HENDRIX
STEVE HACKETT
GEORGE HARRISON
MARTIN PUGH
BALLS
WINGS
MOODY BLUES
BAD COMPANY
BEATLES
PLASTIC ONO BAND
THE EXPERIENCE
GENESIS
STEAMHAMMER
ARMAGEDDON
THE MOVE
GTR


Paul
Ringo
John
George


RONNIE WOOD
JACK BRUCE
MICK TAYLOR
ROD STEWART
KEITH RICHARDS
RAY DAVIES
NICKY HOPKINS
MICK JAGGER
RED BEACH
JOHN PAUL JONES
DICK TAYLOR
MICK AVORY
DAVE DAVIES
ROLLING STONES
JEFF BECK GROUP


Ronnie
MickT
Keith
Mick
Dick

What is a k-club?

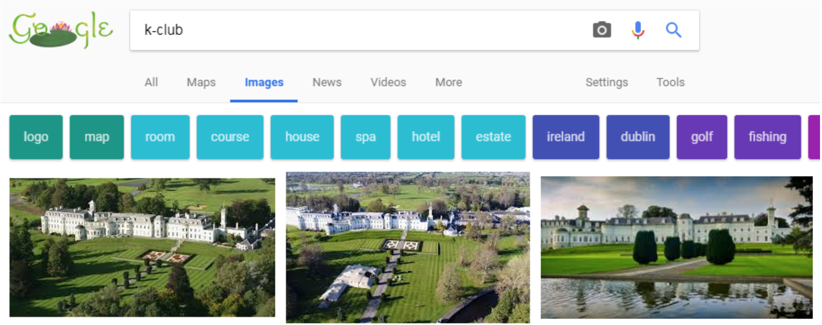
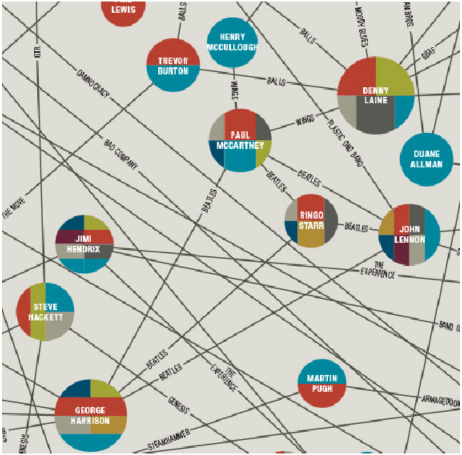
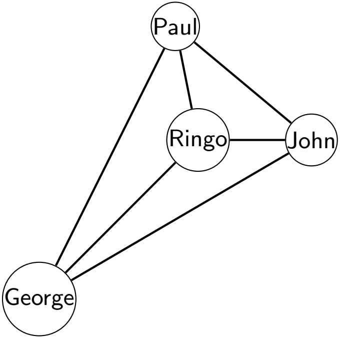


Figure: Did you mean "The K Club Hotel in Straffan, Ireland?"

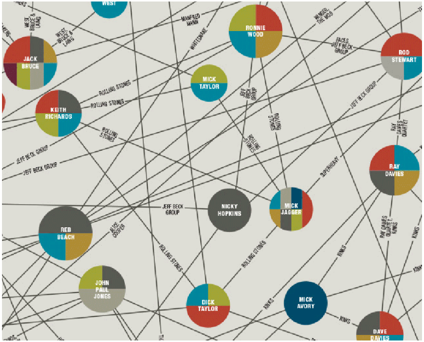
Definition (Mokken, 1979)

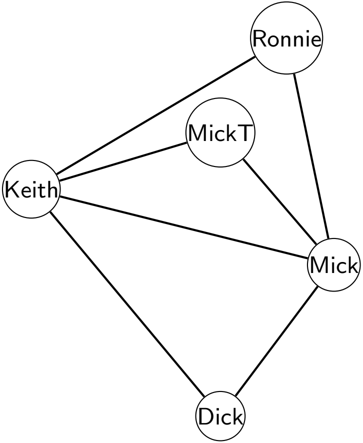
A vertex subset $S \subseteq V$ is a k-club in graph $G = (V, E)$ if $\mathrm{diam}(G[S]) \leq k$.





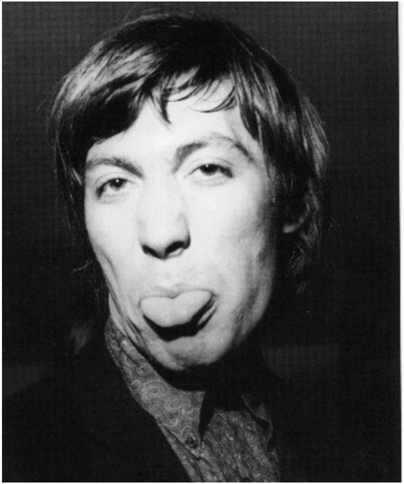
Example 1-club





Example 2-club

"You included Ringo, but not me?"

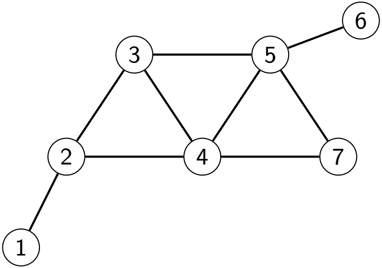


Low-diameter constraints appear in:

- sociology/social network analysis
  - distance k-clique introduced by Alba (Journal of Mathematical Sociology, 1973)
  - k-club by Mokken (Quality & Quantity, 1979)
- bioinformatics
- political districting
- wildlife conservation planning
- wireless sensor networks

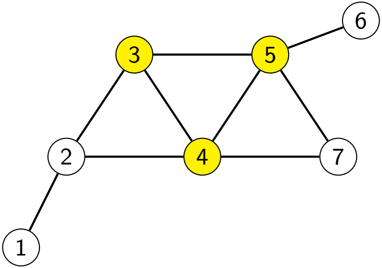
A stylized problem that incorporates these diameter constraints:

Problem: Maximum k-club.
Input: A simple graph $G = (V, E)$ and positive integer k.
Output: A maximum cardinality k-club $S \subseteq V$.



A <u>stylized</u> problem that incorporates these diameter constraints:

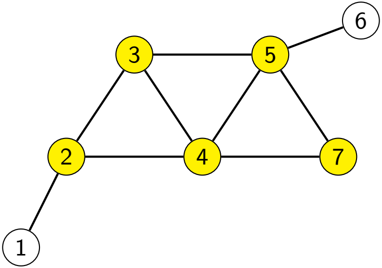
Problem: Maximum k-club.
Input: A simple graph $G = (V, E)$ and positive integer k.
Output: A maximum cardinality k-club $S \subseteq V$.



A maximum 1-club

A stylized problem that incorporates these diameter constraints:

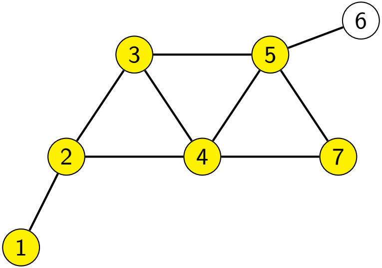
Problem: Maximum k-club.
Input: A simple graph $G = (V, E)$ and positive integer k.
Output: A maximum cardinality k-club $S \subseteq V$.



A maximum 2-club

A stylized problem that incorporates these diameter constraints:

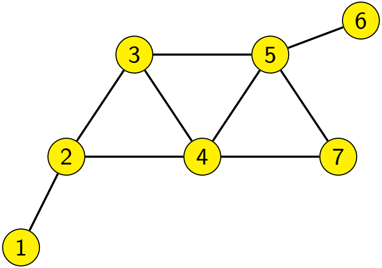
Problem: Maximum k-club.
Input: A simple graph $G = (V, E)$ and positive integer k.
Output: A maximum cardinality k-club $S \subseteq V$.



A maximum 3-club

A stylized problem that incorporates these diameter constraints:

Problem: Maximum k-club.
Input: A simple graph $G = (V, E)$ and positive integer k.
Output: A maximum cardinality k-club $S \subseteq V$.



A maximum 4-club

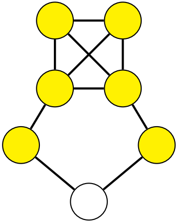
Definition (Alba, 1973)

$S \subseteq V$ is a (distance) k-clique if $\text{dist}_G(i,j) \leq k$ for all $i,j \in S$.

maximum 2-clique

$$\max_{S \subseteq V} |S|$$

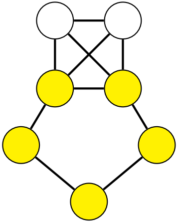
$$\text{dist}_G(i,j) \leq 2 \ \forall i,j \in S.$$



maximum 2-club

$$\max_{S \subseteq V} |S|$$

$$\text{dist}_{G[S]}(i,j) \leq 2 \ \forall i,j \in S.$$



Theorem (Håstad (1999), Zuckerman (2006))

When $k = 1$, hard to get $n^{1-\epsilon}$-approximation for max k-club.

Theorem (Asahiro et al. (2018))

For any fixed $k \geq 2$, there are $n^{1/2}$-approximation algorithms for k-club and k-clique, and getting an $n^{1/2-\epsilon}$-approximation is hard.

Theorem (Mahdavi Pajouh and Balasundaram (2012))

For any fixed $k \geq 2$, k-club maximality testing is coNP-complete.

A history of k-club formulations

The usual variables:

$$x_i = \begin{cases} 1 & \text{vertex } i \text{ is chosen in the } k\text{-club} \\ 0 & \text{otherwise} \end{cases}$$

Some others:

- Variable y_P for each path P of length at most k
  - First described by Bourjolly et al. (EJOR, 2002).
  - Problem: can be $\Omega(n^{k+1})$ variables (for hop-based distances).
  - Improved to $O(n^{k-1})$ by Wotzlaw (arXiv, 2014).
- Variable y_{ij}^t if there exists path of length t from i to j
  - First described by Veremyev and Boginski (EJOR, 2012).
  - Problem: somewhat large at $O(kn^2)$ variables; weak LP.
  - Big M's removed by Veremyev et al. (Networks, 2015).

- Formulations for special k
  - $k = 2$. Common neighborhood formulation:

$$x_a + x_b \leq 1 + x(N(a) \cap N(b))$$

  - $k = 3$. Node and edge variables, by Almeida & Carvalho (Networks, 2012):

$$x_a + x_b \leq 1 + x(N(a) \cap N(b)) + y(E_{ab})$$

    where $E_{ab} = \{\{u, v\} \in E \mid u \in N(a) \setminus N(b),\ \ v \in N(b) \setminus N(a)\}$.

  - $k = 3$. Use n variables and exponentially many constraints (ibid):

$$x_a + x_b \leq 1 + x(N(a) \cap N(b)) + x(S)$$

    where S is a vertex cover of the subgraph induced by E_{ab}.

Our proposed formulations

Path-like formulation:

- Variable y_S for each *minimal* length-k a, b-connector $S \subset V$.
- $O(m^{(k-1)/2})$ variables when k is odd;
- $O(nm^{(k-2)/2})$ variables when k is even;
- generalizes $k = 3$ formulation of Almeida and Carvalho (Networks, 2012)
- improves upon Bourjolly et al. (EJOR, 2002) and Wotzlaw (arXiv, 2014)
- for brevity, we skip this one today

Cut-like formulation:

- Just the x_i variables!
- Constraints have a simple form:

$$x_a + x_b \leq 1 + x(C)$$

- Generalizes $k = 2$ common neighbor formulation
- Generalizes $k = 3$ formulation of Almeida and Carvalho (Networks, 2012)

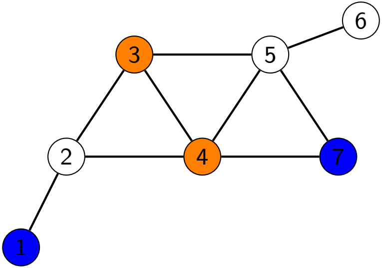
A digression

Vertex-induced connected subgraph polytope:

$$\text{conv.hull}\left\{x^S \in \{0,1\}^n \;\middle|\; G[S] \text{ is connected}\right\}.$$

Practically useful inequalities for a, b-separators $C \subseteq V \setminus \{a, b\}$:

$$x_a + x_b \leq 1 + x(C).$$

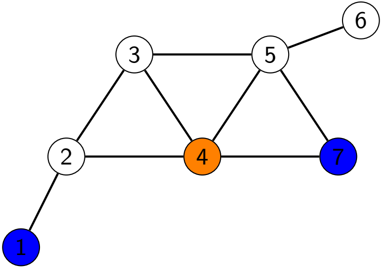


Successfully employed by:

1. Carvajal et al. (OR, 2013) for forest harvest planning
2. Buchanan et al. (IJOC, 2015) for WSNs
3. Fischetti et al. (MPC, 2017) for Steiner problems

See polyhedral study by Wang et al. (MP, 2017).

Is there a distance-constrained generalization?



For 3-club we can write: $x_1 + x_7 \leq 1 + x_4$.

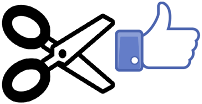
Definition

$C \subseteq V \setminus \{a, b\}$ is a length-k a, b-separator if $\mathrm{dist}_{G-C}(a, b) > k$.

For any length-k a, b-separator C we can write:

$$x_a + x_b \leq 1 + x(C).$$

$$
\begin{aligned}
\max \sum_{i \in V} & x_i \\
& x_a + x_b \leq 1 + x(C) && \forall a, b, C \\
& x_i \in \{0, 1\} && \forall i \in V.
\end{aligned}
$$



Here, $\forall a, b, C$ is shorthand for *minimal* length-k a, b-separators C.

Proposition

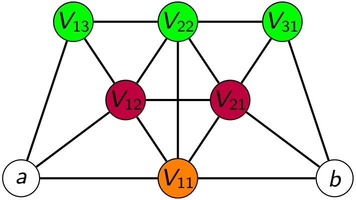
This formulation is correct.

Results skipped today:

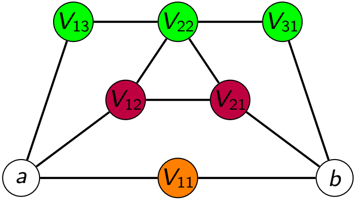
- Formulation strength: CUT $\subsetneq$ PATH
- Facet characterization
- Separation is easy for $k \in \{2, 3, 4\}$ and hard for $k \geq 5$.

Illustration of separation when $k = 4$

$$V_{ij} := \{v \in V \mid \text{dist}(a, v) = i,\ \text{dist}(v, b) = j\}$$



If edge is not in minimal length-4 a, b-connector, ignore it:



Direct arcs to right; node split; weight the new arcs; apply min-cut.

Experiments

PC setup:

- Windows 10; 64bit; VS 2015; C++; 32 GB RAM;
- Intel Xeon Processor E52630 v4:
  - 10 cores (20 threads), 2.2GHz, 3.1GHz Turbo, 25MB cache;
- Gurobi 7.5.1.

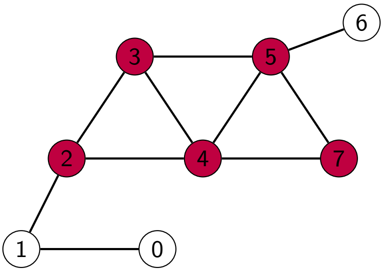
Test instances:

- some real-life graphs from 10th DIMACS Implementation Challenge;
- synthetic graphs developed by Veremyev and Boginski (EJOR, 2012)

The union of these was used by Moradi and Balasundaram (OPTL, 2015).

Heuristic

Might identify this 2-club.

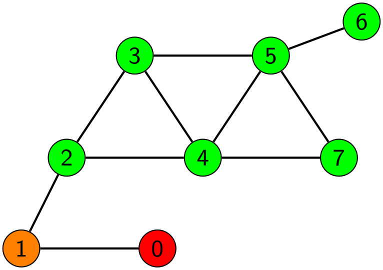


1. initialize $S \leftarrow V$ and create G^k;
2. while S is not a clique in G^k do
   - pick $v \in S$ of minimum degree in $G^k[S]$;
   - $S \leftarrow S \setminus \{v\}$;
3. while S is not a k-club in G do
   - pick $v \in S$ with the fewest nearby nodes $|N^k_{G[S]}(v) \cap S|$ in $G[S]$;
   - $S \leftarrow S \setminus \{v\}$;
4. return S.

A poly-time version of k-clique & DROP by Bourjolly et al. (C&OR, 2000).

Preprocessing

When $k = 2$, red node can be deleted, and then orange node.

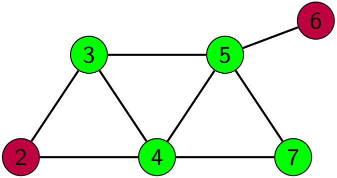


Preprocessing for k-club when given lower bound p:

1. create the k-th power graph G^k;
2. find the $(p-1)$-core $G' = (V', E')$ of G^k;
3. return $G[V']$.

We initialize the formulation with:

$$x_a + x_b \leq 1 \qquad \text{when } \operatorname{dist}_G(a,b) > k$$
$$x_i \in \{0,1\} \qquad i \in V.$$



$x_2 + x_6 \leq 1.$

Add (violated) length-k a,b-separator inequalities on-the-fly.

$$x_a + x_b \leq 1 + x(C).$$

We only separate integral points!

Our separation routine

Given a possible solution $x^* \in \{0,1\}^n$:

1. $D := \{i \in V \mid x_i^* = 1\}$;
2. If D is a k-club, return;
3. For every "far" pair $\{a,b\} \in \binom{D}{2}$ in $G[D]$ do
   - $C := V \setminus D$ is a length-k a,b-separator;
   - Find $C' \subseteq C$ minimal and add $x_a + x_b \leq 1 + x(C')$.

DIMACS 10 Experiments

Times in seconds using the same heuristic, preprocessing, computer, and solver.

		$k=3$				$k=4$			
Graph	n	R	CHC	PATH	CUT	R	CHC	PATH	CUT
karate	34	0.19	0.01	0.02	0.01	0.61	0.01	0.03	0.01
lesmis	77	1.04	0.01	0.04	0.01	2.94	0.01	0.19	0.01
polbooks	105	1.01	0.01	0.07	0.01	13.99	0.03	1.16	0.03
adjnoun	112	6.48	0.03	0.33	0.01	8.90	0.02	1.02	0.02
football	115	[24,63]	[24,67]	9.96	0.27	13.35	0.01	1.35	0.01
celegansm	453	337.12	0.06	5.41	0.07	207.38	0.07	39.14	0.08
email	1133	LPNS	[192,232]	[206,239]	264.41	LPNS	[642,653]	LPNS	1.88
polblogs	1490	LPNS	2.37	960.54	1.56	LPNS	0.78	LPNS	0.72
netscience	1589	1.30	0.03	0.07	0.03	2.38	0.03	0.23	0.03
power	4941	1.21	0.95	1.00	0.99	1.77	0.98	1.41	0.93
hep-th	8361	[114,124]	[114,122]	2622.90	191.30	LPNS	[318,346]	[336,347]	428.68
PGPgiantc	10680	633.09	6.08	25.25	5.54	1556.62	6.38	407.25	5.81

- R – Recursive, Veremyev et al. (Networks, 2015)
- CHC – Canonical hypercube cut, Moradi and Balasundaram (OPTL, 2015)

Synthetic Graph Experiments

Times reported are averages over 10 instances. If not all 10 solved in 1 hour, the number solved is reported in parenthesis.

		$k=3$				$k=4$			
n	ρ (%)	R	CHC	PATH	CUT	R	CHC	PATH	CUT
100	2	0.56	0.02	0.05	0.02	2.56	0.02	0.15	0.02
	3	6.18	0.14	0.25	0.09	52.42	(7)	1.29	0.13
	4	41.34	(8)	1.33	0.29	550.42	(1)	1.88	0.11
200	1	2.54	0.12	0.27	0.12	6.29	0.14	0.72	0.16
	1.5	16.42	0.70	1.35	0.75	285.07	34.95	6.04	0.70
	2	201.70	2.60	8.00	2.09	(0)	(0)	551.84	21.07
300	0.5	0.13	0.01	0.02	0.01	1.17	0.04	0.10	0.04
	1	33.94	1.88	4.96	1.95	235.45	1.85	10.77	2.03
	1.5	444.22	7.62	23.31	6.60	(0)	(1)	(7)	452.34

PATH, which has many variables, is omitted for larger k.

		$k=5$			$k=6$			$k=7$		
n	ρ (%)	R	CHC	CUT	R	CHC	CUT	R	CHC	CUT
100	2	5.29	(9)	0.02	19.37	0.51	0.02	4.69	0.03	0.01
	3	22.15	(7)	0.03	10.24	0.11	0.01	9.54	0.01	0.01
	4	8.77	0.71	0.01	9.53	0.01	0.01	10.76	0.01	0.01
200	1	86.08	1.19	0.28	(9)	(5)	0.23	(8)	(5)	0.12
	1.5	(2)	(0)	1.02	(4)	(1)	0.12	153.08	(8)	0.04
	2	(4)	(0)	0.17	46.08	0.03	0.02	59.00	0.02	0.02
300	0.5	1.46	0.04	0.04	7.68	0.04	0.05	9.67	0.05	0.05
	1	(0)	(0)	6.72	(0)	(0)	2.23	(6)	(2)	0.14
	1.5	(0)	(0)	1.41	(7)	(9)	0.04	(6)	0.04	0.03

Conclusion

New formulations for low-diameter clusters (k-clubs):

- path-like and cut-like formulations
- generalize or strengthen several previous formulations
- cut-like also applies when distances are not hop-based

Top performer is cut-like formulation:

- uses n variables
- usually solves in 1 or 2 seconds
- all 24 real-life instances in < 8 minutes (others fail on 4+)
- all 450 synthetic instances in < 35 minutes (others fail on 84+)

Manuscript available on Optimization Online:

- Austin Buchanan and Hosseinali Salemi. Parsimonious formulations for low-diameter clusters. `http://www.optimization-online.org/DB_HTML/2017/09/6196.html`

References, part I.

- R.D. Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126, 1973.
- M.T. Almeida, F.D. Carvalho. Integer models and upper bounds for the 3-club problem. *Networks*, 60(3): 155–166, 2012.
- Y. Asahiro, Y. Doi, E. Miyano, H. Shimizu. Optimal approximation algorithms for maximum distance-bounded subgraph problems.I n: Lu Z., Kim D., Wu W., Li W., Du DZ. (eds) Combinatorial Optimization and Applications. Lecture Notes in Computer Science, vol 9486, 586–600, 2015.
- G. Baier, T. Erlebach, A. Hall, E. Köhler, P. Kolman, O. Pangrác, H. Schilling, M. Skutella. Length-bounded cuts and flows. *ACM Transactions on Algorithms (TALG)*, 7(1):4, 2010.
- J.-M. Bourjolly, G. Laporte, G. Pesant. Heuristics for finding k-clubs in an undirected graph. *Computers & Operations Research*, 27(6):559–569, 2000.
- J.-M. Bourjolly, G. Laporte, G. Pesant. An exact algorithm for the maximum k-club problem in an undirected graph. *European Journal of Operational Research*, 138(1): 21–28, 2002.
- A. Buchanan, J.S. Sung, S. Butenko, E.L. Pasiliao. An integer programming approach for fault-tolerant connected dominating sets. *INFORMS Journal on Computing*, 27(1):178–188, 2015.
- A. Buchanan, H. Validi. The optimal design of low-latency virtual backbones. Revision submitted.
- R. Carvajal, M. Constantino, M. Goycoolea, J.P. Vielma, and A. Weintraub. Imposing connectivity constraints in forest planning models. *Operations Research*, 61(4):824–836, 2013.
- M. Fischetti, M. Leitner, I. Ljubić, M. Luipersbeck, M. Monaci, M. Resch, D. Salvagnin, M. Sinnl. Thinning out Steiner trees: a node-based model for uniform edge costs. *Mathematical Programming Computation*, 9(2):203–229, 2017.

References, part II.

- L. Lovász, V. Neumann-Lara, M. Plummer. Mengerian theorems for paths of bounded length. *Periodica Mathematica Hungarica*, 9(4):269–276, 1978.
- F. Mahdavi Pajouh, B. Balasundaram, I. Hicks. On the 2-club polytope of graphs. *Operations Research*, 64(5):1466–1481, 2016.
- F. Mahdavi Pajouh, B. Balasundaram. On inclusionwise maximal and maximum cardinality k-clubs in graphs. *Discrete Optimization*, 9(2):84–97, 2012.
- A.R. Mahjoub, S.T. McCormick. Max flow and min cut with bounded-length paths: complexity algorithms, and approximation. *Mathematical Programming*, 124(1):271–284, 2010.
- D.W. Matula, L.L. Beck. Smallest-last ordering and clustering and graph coloring algorithms. *Journal of the ACM (JACM)*, 30(3):417–427, 1983.
- R.J. Mokken. Cliques, clubs and clans. *Quality & Quantity*, 13(2):161–173, 1979.
- E. Moradi, B. Balasundaram. Finding a maximum k-club using the k-clique formulation and canonical hypercube cuts. *Optimization Letters*, 2015. DOI 10.1007/s11590-015-0971-7
- J. Pattillo, N. Youssef, S. Butenko. On clique relaxation models in network analysis. *European Journal of Operational Research*, 226(1):9–18, 2013.
- A. Veremyev, V. Boginski. Identifying large robust network clusters via new compact formulations of maximum k-club problems. *European Journal of Operational Research*, 218(2), 316–326, 2012.
- A. Veremyev, O.A. Prokopyev, E.L. Pasiliao. Critical nodes for distance-based connectivity and related problems in graphs. *Networks*, 66(3):170–195, 2015.
- J.L. Walteros and A. Buchanan. Why is maximum clique easy in practice? Submitted.
- Y. Wang, A. Buchanan, S. Butenko. On imposing connectivity constraints in integer programs. *Mathematical Programming*, 166(1): 241–271, 2017.
- A. Wotzlaw. On solving the maximum k-club problem. arXiv:1403.5111v2, 2014.