Charalampos (Babis) E. Tsourakakis
charalampos.tsourakakis@aalto.fi

Denser than the Densest Subgraph: Extracting Optimal Quasi-Cliques with Quality Guarantees

KDD 2013

Francesco Bonchi
Yahoo! Research

Aristides Gionis
Aalto University

Francesco Gullo
Yahoo! Research

Maria Tsiarli
University of Pittsburgh

Denser than the densest

- Densest subgraph problem is very popular in practice. However, not what we want for many applications.
- δ=edge density,D=diameter,τ=triangle density

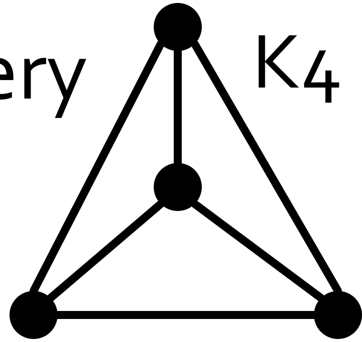
	densest subgraph				optimal quasi-clique			
	$\frac{\lvert S\rvert}{\lvert V\rvert}$	δ	D	τ	$\frac{\lvert S\rvert}{\lvert V\rvert}$	δ	D	τ
Dolphins	0.32	0.33	3	0.04	0.12	0.68	2	0.32
Football	1	0.09	4	0.03	0.10	0.73	2	0.34
Jazz	0.50	0.34	3	0.08	0.15	1	1	1
Celeg. N.	0.46	0.13	3	0.05	0.07	0.61	2	0.26

Graph mining applications

- Thematic communities and spam link farms [Gibson, Kumar, Tomkins '05]
- Graph visualization[Alvarez-Hamelin etal.'05]
- Real time story identification [Angel et al. '12]
- Motif detection [Batzoglou Lab '06]
- Epilepsy prediction [Iasemidis et al. '01]
- Finding correlated genes [Horvath et al.]
- Many more ..

Measures

- Clique: each vertex in S connects to every other vertex in S.



- α-Quasi-clique:
  the set S has at least $\alpha|S|(|S|-1)/2$ edges.

- k-core: every vertex connects to at least k other vertices in S.

Measures

- $\delta(S) = \frac{e[S]}{\binom{|S|}{2}}$ *Density*
- $d(S) = \frac{2e[S]}{|S|}$ *Average degree*
- $t(S) = \frac{t[S]}{\binom{|S|}{3}}$ *Triangle Density*

Contributions

- General framework which subsumes popular density functions.
- Optimal quasi-cliques.
- An algorithm with additive error guarantees and a local-search heuristic.
- Variants
  - Top-k optimal quasi-cliques
  - Successful team formation

Contributions

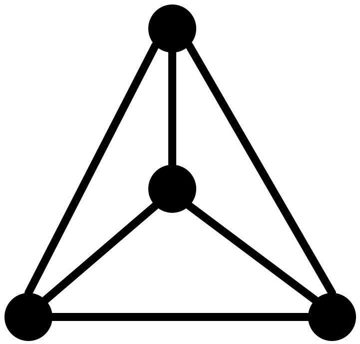
- Experimental evaluation
  - Synthetic graphs
  - Real graphs

- Applications
  - Successful team formation of computer scientists
  - Highly-correlated genes from microarray datasets

First, some related work.

Cliques

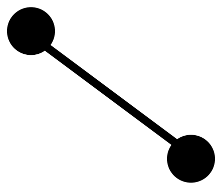
- Maximum clique problem:
  find clique of maximum possible size.
  NP-complete problem
- Unless P=NP, there cannot be a polynomial time algorithm that approximates the maximum clique problem within a factor better than $O(n^{1-\varepsilon})$ for any ε>0 [Håstad '99].



K4

(Some) Density Functions

- $\delta(S) = \frac{e[S]}{\binom{|S|}{2}}$ A single edge achieves always maximum possible $\delta(S)$



- $d(S) = \frac{e[S]}{|S|}$ Densest subgraph problem
- $d(S) = \frac{e[S]}{|S|}, |S| = k$ k-Densest subgraph problem
- $d(S) = \frac{e[S]}{|S|}, |S| \geq k \ (|S| \leq k)$ DalkS (Damks)

Densest Subgraph Problem

- Maximize average degree
- Solvable in polynomial time
  - Max flows (Goldberg)
  - LP relaxation (Charikar)
- Fast ½-approximation algorithm (Charikar)

k-Densest subgraph

- k-densest subgraph problem is NP-hard
- Feige, Kortsatz, Peleg
  Bhaskara, Charikar, Chlamtac, Vijayraghavan
  Asahiro et al.
  Andersen
  Khuller, Saha [approximation algorithms],
  Khot [no PTAS].

Quasicliques

- A set S of vertices is α-quasiclique if

$$e[S] \geq \alpha\binom{|S|}{2}$$

- [Uno '10] introduces an algorithm to enumerate all α-quasicliques.

Edge-Surplus Framework

- For a set of vertices S define

$$f_\alpha(S) = g(e[S]) - ah(|S|)$$

where g,h are both strictly increasing, α>0.

- Optimal (α,g,h)-edge-surplus problem
  Find S* such that $f_\alpha(S^*) \geq f_\alpha(S)$.

Edge-Surplus Framework

- When g(x)=h(x)=log(x), α=1, then

Optimal (α,g,h)-edge-surplus problem becomes $\max \log \frac{e[S]}{|S|}$, which is the densest subgraph problem.

- g(x)=x, h(x)=0 if x=k, o/w +∞ we get the k-densest subgraph problem.

Edge-Surplus Framework

- When g(x)=x, h(x)=x(x-1)/2 then we obtain $\max_{S \subseteq V, |S| \geq 2} e[S] - \alpha\binom{|S|}{2}$, which we define as the optimal quasiclique (OQC) problem.
- Theorem 1: Let g(x)=x, h(x) concave. Then the optimal (α,g,h)-edge-surplus problem is poly-time solvable.
  - However, this family is not well suited for applications as it returns most of the graph.

Hardness of OQC

- Conjecture: finding a planted clique C of size $n^{\frac{1}{2}-\delta}, \delta > 0$ in a random binomial graph $G\left(n, \frac{1}{2}\right)$ is hard.

- Let $f(S) = e[S] - \frac{2}{3}\binom{|S|}{2}$. Then,
  $f(C) = \frac{1}{3}\binom{n^{\frac{1}{2}-\delta}}{2} > 0,$
  $E[f(S)] = -\frac{1}{6}\binom{|S|}{2} < 0.$

Multiplicative approximation algorithms

- Notice that in general the optimal value can be negative.
- We can obtain guarantees for a shifted objective but introduces large additive error making the algorithm *almost* useless, i.e., *except for very special graphs*.
- Other type of guarantees more suitable.

Optimal Quasicliques

- Additive error approximation algorithm
  - $G_n \leftarrow G$
  - For $k \leftarrow n$ downto 1
    - Let v be the smallest degree vertex in G_k.
    - $G_{k-1} \leftarrow G_k - \{v\}$
  - Output $\bar{S} \leftarrow argmax_{1 \le k \le n}\, f_a(G_k)$

Theorem: $f_\alpha(\bar{S}) \ge f_\alpha(S^*) - \frac{\alpha}{2}\text{"}small\text{"} \times |\bar{S}|$

Running time: O(n+m). However it would be nice to have running time O(|output|).

Optimal Quasicliques
Local Search Heuristic

1. Initialize S with a random vertex.
2. For t=1 to T_{max}
   1. Keep expanding S by adding at each time a vertex $v \notin S$ such that $f_\alpha(S \cup v) \geq f_\alpha(S)$.
   2. If not possible see whether there exist $v \in S$ such that $f_\alpha(S - \{v\}) \geq f_\alpha(S)$.
      1. If yes, remove it. Go back to previous step.
      2. If not, stop and output S.

Experiments

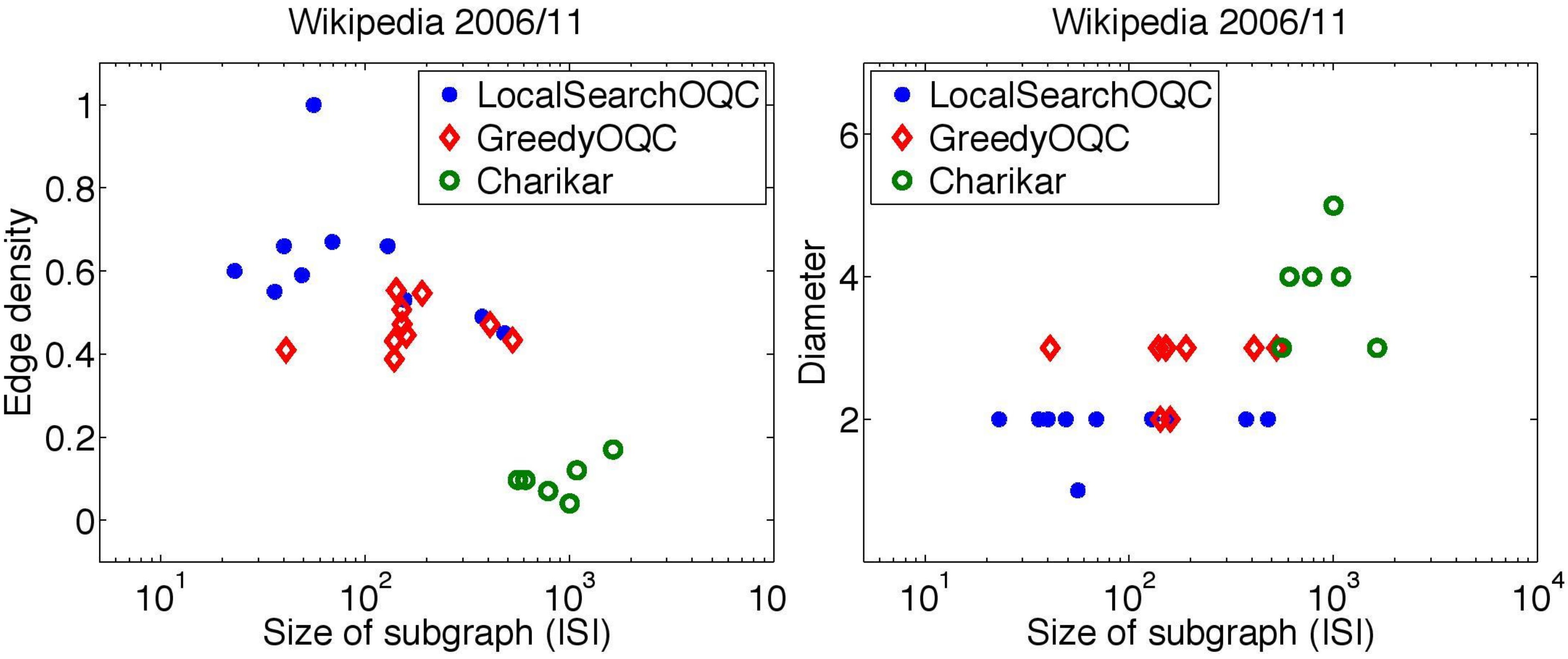
	Vertices	Edges	Description
Dolphins	62	159	Biological Network
Polbooks	105	441	Books Network
Adjnoun	112	425	Adj. and Nouns in 'David Copperfield'
Football	115	613	Games Network
Jazz	198	2 742	Musicians Network
Celegans N.	297	2 148	Biological Network
Celegans M.	453	2 025	Biological Network
Email	1 133	5 451	Email Network
AS-22july06	22 963	48 436	Auton. Systems
Web-Google	875 713	3 852 985	Web Graph
Youtube	1 157 822	2 990 442	Social Network
AS-Skitter	1 696 415	11 095 298	Auton. Systems
Wikipedia 2005	1 634 989	18 540 589	Web Graph
Wikipedia 2006/9	2 983 494	35 048 115	Web Graph
Wikipedia 2006/11	3 148 440	37 043 456	Web Graph

Experiments

	$\lvert S \rvert$			δ			D			τ		
	densest subgraph	opt. quasi-clique		densest subgraph	opt. quasi-clique		densest subgraph	opt. quasi-clique		densest subgraph	opt. quasi-clique	
		GREEDY	LS		GREEDY	LS		GREEDY	LS		GREEDY	LS
Dolphins	19	13	8	0.27	0.47	0.68	3	3	2	0.05	0.12	0.32
Polbooks	53	13	16	0.18	0.67	0.61	6	2	2	0.02	0.28	0.24
Adjnoun	45	16	15	0.20	0.48	0.60	3	3	2	0.01	0.10	0.12
Football	115	10	12	0.09	0.89	0.73	4	2	2	0.03	0.67	0.34
Jazz	99	59	30	0.35	0.54	1	3	2	1	0.08	0.23	1
Celeg. N.	126	27	21	0.14	0.55	0.61	3	2	2	0.07	0.20	0.26
Celeg. M.	44	22	17	0.35	0.61	0.67	3	2	2	0.07	0.26	0.33
Email	289	12	8	0.05	1	0.71	4	1	2	0.01	1	0.30
AS-22july06	204	73	12	0.40	0.53	0.58	3	2	2	0.09	0.19	0.20
Web-Google	230	46	20	0.22	1	0.98	3	2	2	0.03	0.99	0.95
Youtube	1874	124	119	0.05	0.46	0.49	4	2	2	0.02	0.12	0.14
AS-Skitter	433	319	96	0.41	0.53	0.49	2	2	2	0.10	0.19	0.13
Wiki '05	24555	451	321	0.26	0.43	0.48	3	3	2	0.02	0.06	0.10
Wiki '06/9	1594	526	376	0.17	0.43	0.49	3	3	2	0.10	0.06	0.11
Wiki '06/11	1638	527	46	0.17	0.43	0.56	3	3	2	0.31	0.06	0.35

	DS	M1	M2	DS	M1	M2	DS	M1	M2	DS	M1	M2
Wiki '05	24.5 K	451	321	.26	.43	.48	3	3	2	.02	.06	.11
Youtube	1.9K	124	119	0.05	0.46	0.49	4	2	2	.02	.12	.14

Top-k densest subgraphs



Constrained Optimal Quasicliques

- Given a set of vertices Q

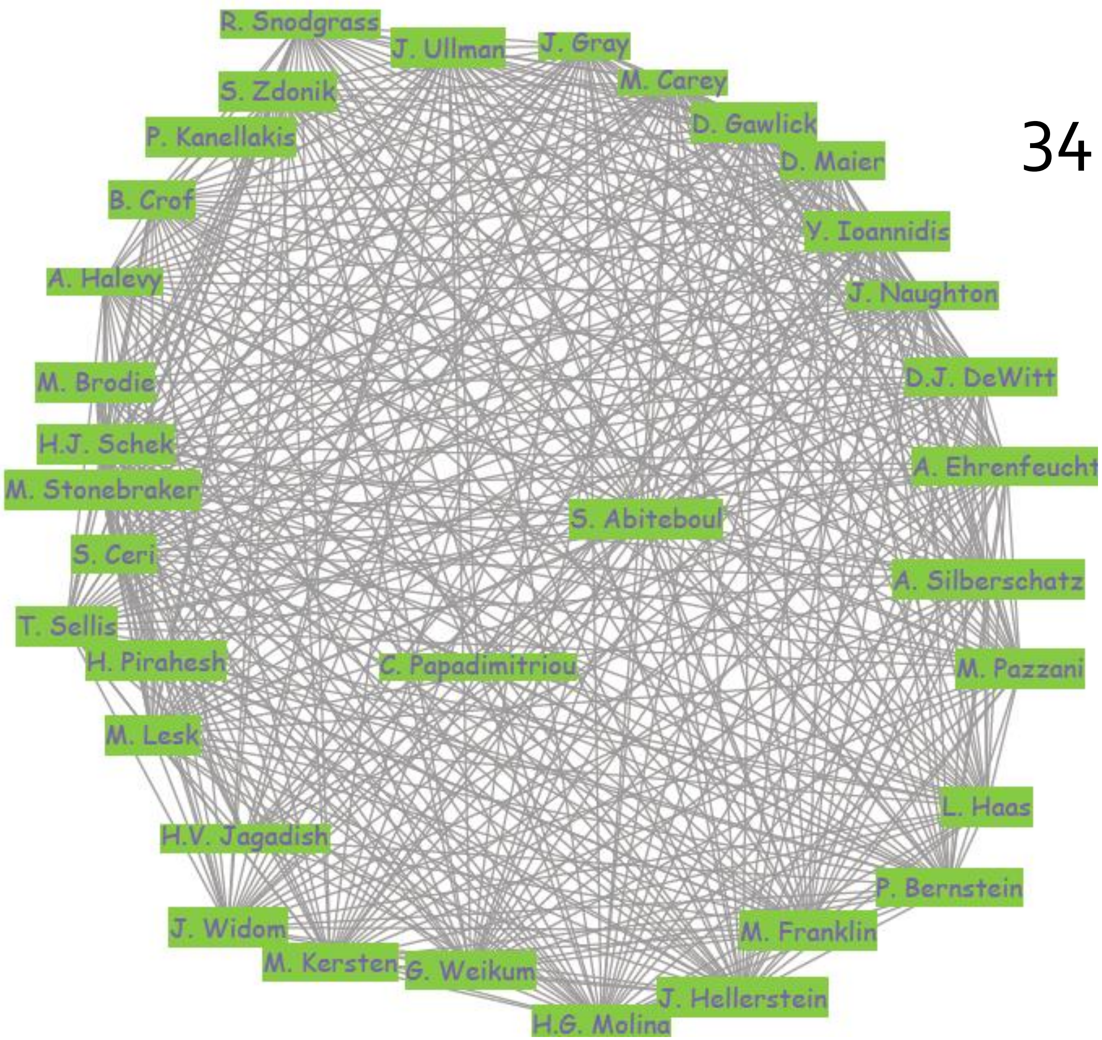
$$\max_{Q \subseteq S \subseteq V} f_\alpha(S) = \max_{Q \subseteq S \subseteq V} e[S] - \alpha \binom{|S|}{2}$$

- Lemma: NP-hard problem.
- Observation: Easy to adapt our efficient algorithms to this setting.
  - Local Search: Initialize S with Q and never remove a vertex if it belongs to Q
  - Greedy: Never peel off a vertex from Q

Application 1

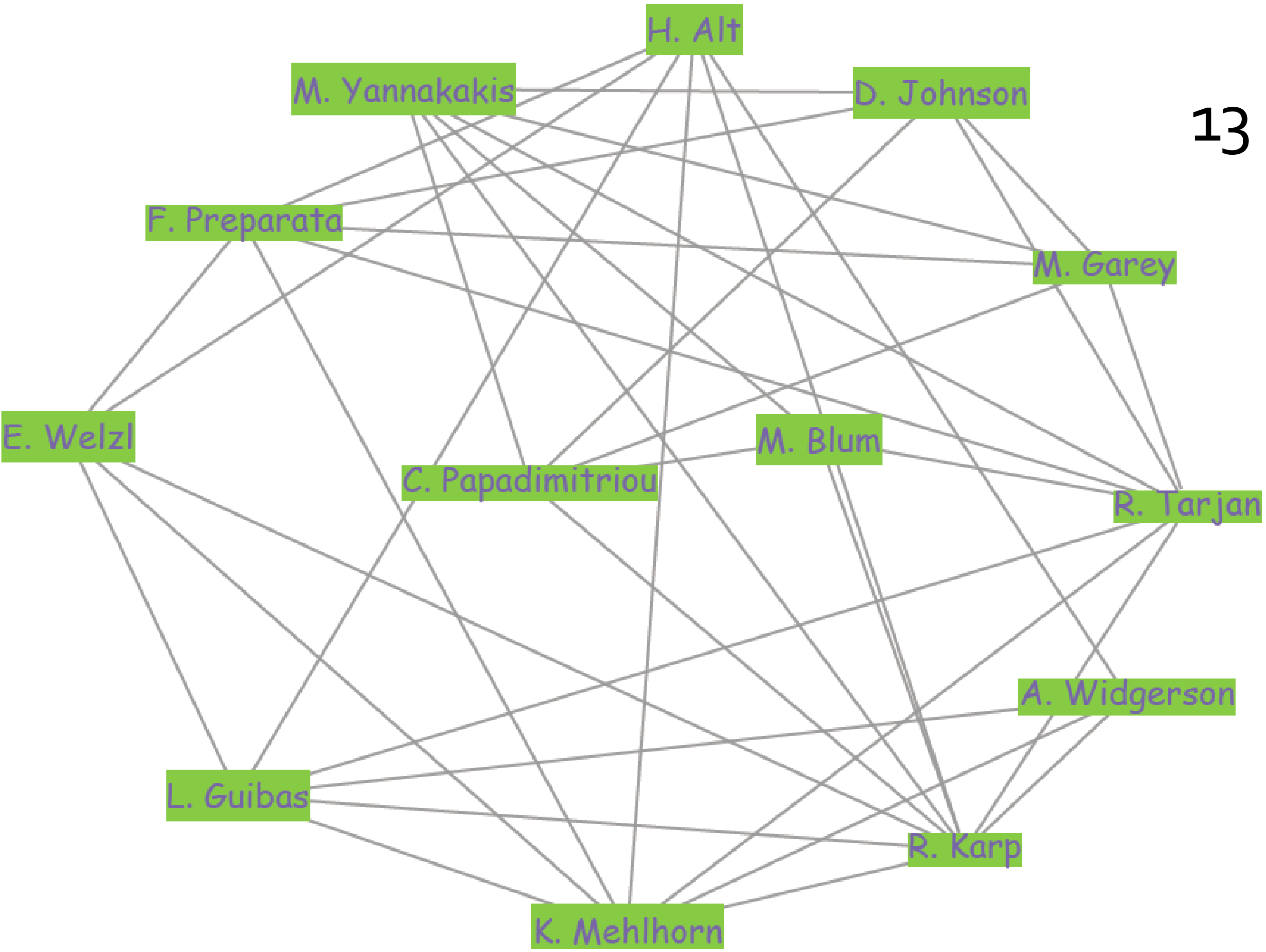
- Suppose that a set Q of scientists wants to organize a workshop. How do they invite other scientists to participate in the workshop so that the set of all participants, including Q, have similar interests ?

Query 1, Papadimitriou and Abiteboul



34 vertices , $\delta(S)$= 0.81

Query 2, Papadimitriou and Blum

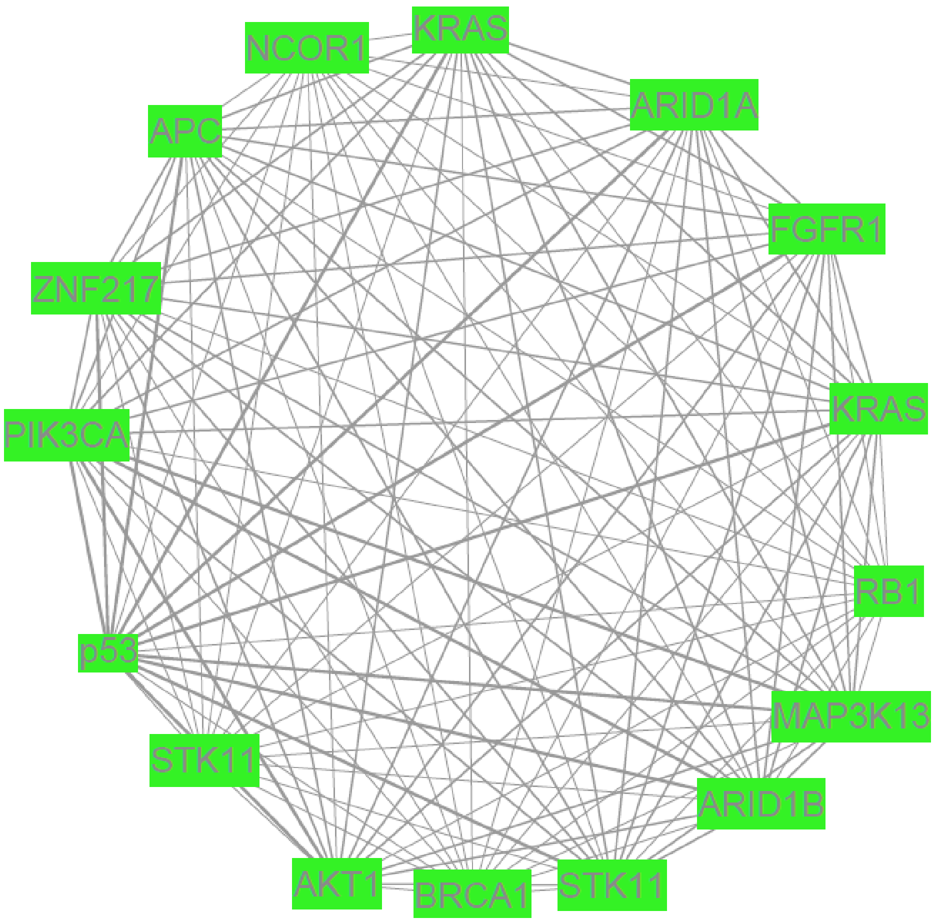


13 vertices, $\delta(S)$=0.49

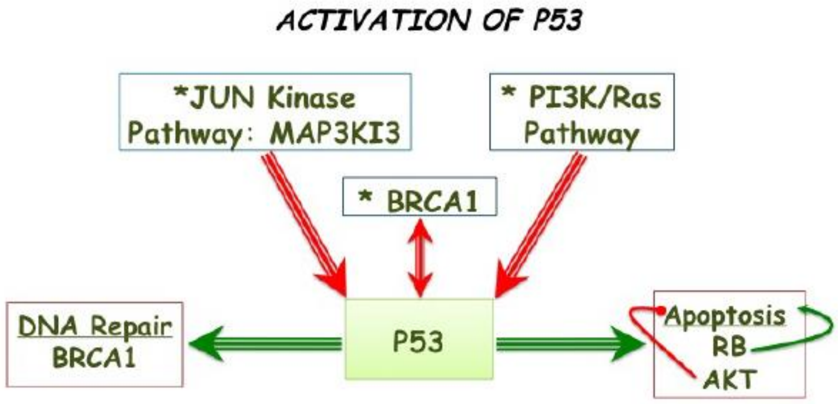
Application 2

- Given a microarray dataset and a set of genes Q, find a set of genes S that includes Q and they are all highly correlated.
- Co-expression network
  - Measure gene expression across multiple samples
  - Create correlation matrix
  - Edges between genes if their correlation is > ρ.
- A dense subgraph in a co-expression network corresponds to a set of highly correlated genes.

Query, p53



NCOR1
KRAS
ARID1A
APC
FGFR1
ZNF217
PIK3CA
KRAS
RB1
p53
MAP3K13
STK11
ARID1B
AKT1
BRCA1
STK11



ACTIVATION OF P53
*JUN Kinase Pathway: MAP3KI3
* PI3K/Ras Pathway
* BRCA1
DNA Repair BRCA1
P53
Apoptosis RB AKT

Future Work

- Hardness
- Analysis of local search algorithm
- Other algorithms with additive approximation guarantees
- Study the natural family of objectives

$$\max_{S \subseteq V, |S| \geq 2} e[S] - \alpha |S|^{\gamma}, \gamma > 1$$

Thank you!