Finding Dense and Connected Subgraphs in Dual Networks

Zohreh Raghebi

Spring 2017

BIG Data Management and Mining Laboratory
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

Authors

- ICDE 2015

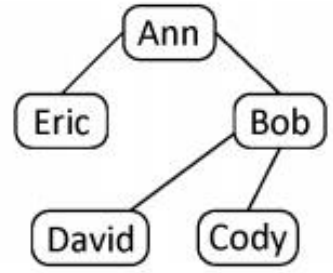
Yubao Wu[1], Ruoming Jin[2], Xiaofeng Zhu[3], Xiang Zhang[1]
[1]*Department of Electrical Engineering and Computer Science, Case Western Reserve University,*
[2]*Department of Computer Science, Kent State University,*
[3]*Department of Epidemiology and Biostatistics, Case Western Reserve University,*
[1]{*yubao.wu, xiang.zhang*}*@case.edu,* [2]*jin@cs.kent.edu,* [3]*xxz10@case.edu*

Problem Definition

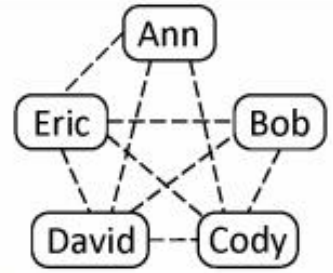
- Finding the densest subgraph is an important problem
- Most of the existing work focuses on a single network
- Given a graph $G(V, E)$, find the subgraph with maximum density (average edge weight)
- In many real-life applications, we often observe two complementary networks:
  - The *physical* interaction among a set of nodes
  - The *conceptual* interaction

Motivation

- A conceptual user similarity network can be derived by measuring the correlation of common ratings between two users
- Two users with similar interests may not have direct physical contact
- However, if a set of users with highly similar interest
- Also physically connected, it may be utilized for recommendation
- If one of the users receives an advertisement of an interested product
  - This information is likely to propagate to the rest of the group
- Because of their common interest and physical connectivity



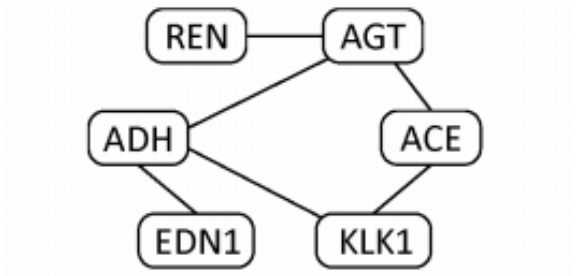
(a) Physical connectivity network



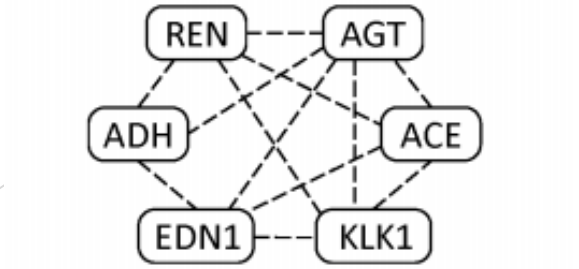
(b) Interest similarity network

Motivation

- In such applications, it is important to find subgraphs that are dense in the conceptual network
- Also connected in the physical network
- In genetics, it is crucial to interpret genetic interactions by protein interactions.
- The genetic interaction network represents the conceptual interaction among genes,
  - Where the interactions are measured by statistical test
- Two genes with strong genetic interaction may not have physical interaction
- The protein interaction network represents physical interactions



(a) Protein interaction network

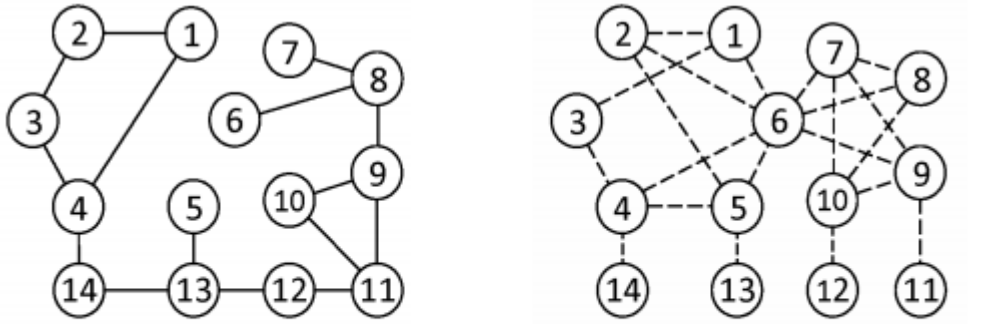


(b) Genetic interaction network

Problem Definition

- Given two graphs *Ga*(*V,Ea*) and *Gb*(*V, Eb*) representing the physical and conceptual networks respectively, the DCS consists of a subset of nodes $S \subseteq V$ such that the induced subgraph *Ga*[*S*] is connected and the density of *Gb*[*S*] is maximized

- Variations of the densest subgraph problem:
- The densest *k* subgraph problem aims to find the densest subgraph with exactly *k* nodes
- The problem of finding the densest subgraph with seed nodes requires that a set of input nodes must be included in the resulting subgraph



(a) Physical network G_a (b) Conceptual network G_b

Fig. 4. An example of dual networks

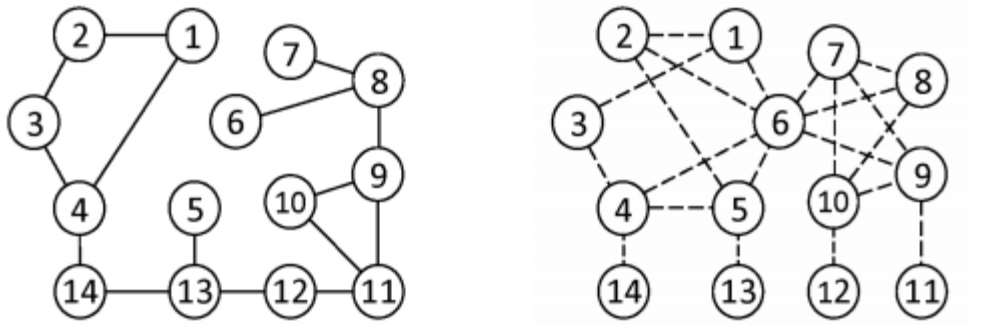
Approaches

- A two-step approach to solve this problem
- In the first step, we show that by removing low degree leaf nodes in the dual network
  - We can not only dramatically reduce the search space,
- In the second step, we develop two greedy approaches to find the DCS in the pruned dual networks
- The first approach finds the densest subgraph in the conceptual network first
- Then refines and makes it connected in the physical network.

- The second approach keeps the target subgraph connected in the physical network while deleting low degree nodes in the conceptual network.

Pruning Stage

- To remove the *low degree leaf nodes* from the dual networks
- Guarantees that the optimal DCS is contained in the resulting networks.
- Low Degree Leaf nodes:
- Given dual networks $G(V,Ea,Eb)$, suppose that its DCS consists of a set of nodes S.
- Let $\rho(S)$ represent its density in Gb.
- A node $u \in V$ is a low degree leaf node if (1) u is a leaf node in Ga and (2) its degree in Gb is less than $\rho(S)$, $wGb(u) < \rho(S)$
- Lemma:The DCS in dual networks does not contain any low degree leaf node



(a) Physical network G_a (b) Conceptual network G_b

Fig. 4. An example of dual networks

Prunning steps

- Let $G0 = G$ be the original dual networks
- We remove all low degree leaf nodes (using density $\rho(Gb[V])$) in the physical network $Ga0$ and conceptual network$Gb0$
- That is, we remove all the nodes that have degree one in Ga and have degree less than $\rho(Gb[V])$ in Gb from the dual networks.
- Let the resulting dual networks be $G1(V1, Ea(V1), Eb(V1))$. We then continue to remove the low degree leaf nodes using density $\rho(V1)$ in $G1$.
- That is, we remove all the nodes that have degree one in $Ga[V1]$ and have degree less than $\rho(Gb[V1])$ in Gb from the dual networks.
- We repeat this process until no such nodes left

Using this pruning strategy, we can safely remove the nodes that are not in the DCS, thus reduce the overall search space

Greedy Algorithms

- The DCS RDS algorithm first finds the densest subgraph in *Gb*, which usually is disconnected in *Ga*
- It then refines the subgraph by connecting its disconnected components in *Ga*
- Although the densest subgraph can be identified in polynomial time by the parametric maximum flow method
- Its actual complexity $O(nm \log(n2/m))$ is prohibitive for large graphs
- First introduce an effective procedure that can dramatically reduce the cost of finding the densest subgraph in a single graph

Fast Densest Subgraph Finding in Conceptual Network

- Our node removal procedure is based on the following key observation

Lemma 2: Let $\rho(T)$ be the density of the densest subgraph $G[T]$. Any node $u \in T$ has degree $w_{G[T]}(u) \geq \rho(T)$.

- The *d*-core *D* of *G* is the maximal subgraph of *G* such that for any node *u* in *D*, *wD*(*u*) ≥ *d*.

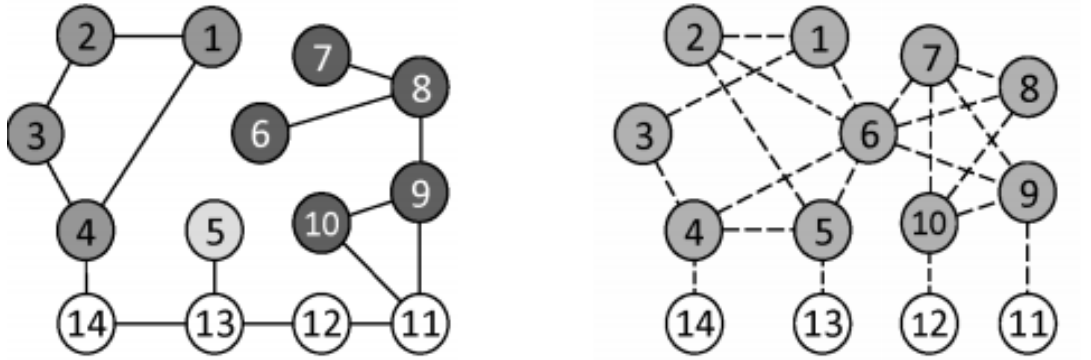
  Note that the *d*-core of a graph is unique and may consist of multiple connected components

Fast Densest Subgraph

- To sum up, we use the following three-step procedure to find the exact densest subgraph from *G*
- (1) Find approximate densest subgraph in *G*, where the density of the discovered subgraph *d*
- (2) Find the *d*-core *D* of *G*
- (3) Compute the exact densest subgraph from *D*

Refining Subgraph in Physical Network

Example 2: Continue the example in Figure 6. The densities of the connected components in the physical network are $\rho(V_1 = \{6, 7, 8, 9, 10\}) = 1.6$, $\rho(V_2 = \{1, 2, 3, 4\}) = 0.75$, and $\rho(V_3 = \{5\}) = 0$. Initially, the subgraph induced by $S_1 = V_1$ has density $\rho(S_1) = 1.6$. Algorithm 1 first connects S_1 and V_2 through the shortest path $H_1 = \{11, 12, 13, 14\}$. The subgraph induced by $S_2 = S_1 \cup V_2 \cup H_1$ has density $\rho(S_2) = 1.31$. After merging V_3, the subgraph induced by S_3 has density $\rho(S_3) = 1.5$. Therefore, the subgraph induced by S_1 has the largest density in G_b and is returned as the DCS.



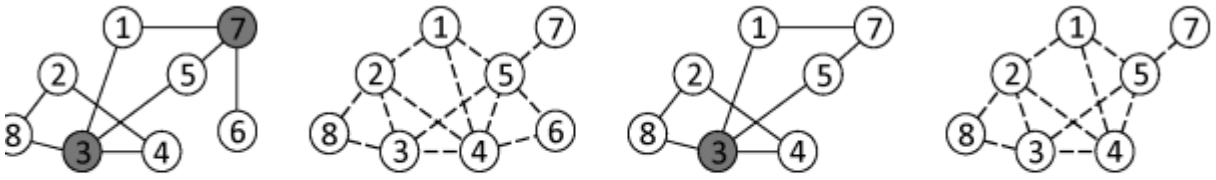
(a) Physical network G_a (b) Conceptual network G_b

Fig. 6. Refining the densest subgraph

THE DCS GND ALGORITHM

- The basic DCS GND algorithm keeps deleting nodes with low degree in the conceptual network
- While avoiding disconnecting the physical network
- A node is an articulation node if removing this node and the edges incident to it disconnects the graph

  Delete one node in each iteration
- The deleted node has the minimum degree in the conceptual network among all the non-articulation nodes in the physical network
- Density of the subgraphs generated in this process is recorded and the subgraph with the largest density is returned as the identified DCS



(a) G_0^a (b) G_0^b (c) G_1^a (d) G_1^b

Fig. 7. Greedy node deletion example

Experiment results

- *Real Dual Networks*: We use the DBLP dataset to build two dual networks
- one for data mining research community and one for database research community
- To construct the dual networks for the data mining community, we extract a set of papers published in 5 data mining conferences: KDD, ICDM, SDM, PKDD and CIKM.
- The dataset contains 4,284 papers and 7,169 authors.
- The physical network is the co-author network with authors being the nodes and edges representing two authors have co-authored a paper

  The conceptual research interest similarity network among authors is constructed based on the similarity of the terms in the paper titles of different authors.
- The shrunk Pearson correlation coefficient is used to compute the research interest similarity between authors

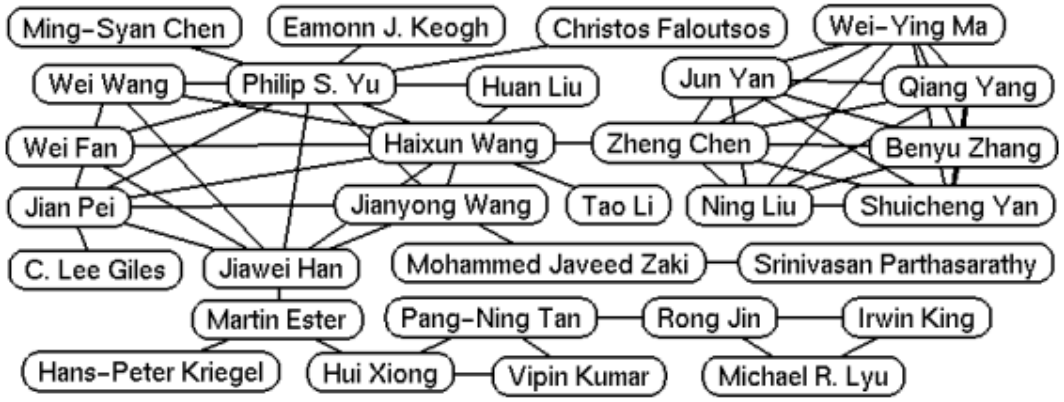
Real datasets: dual networks

- We construct two dual networks using recommender system datasets, Flixster and Epinions
- In the original Flixster dataset, the physical network has 786,936 nodes (users) and 7,058,819 edges representing their social connectivity
- We construct the conceptual interest similarity network by measuring the correlation coefficients of the common ratings between users

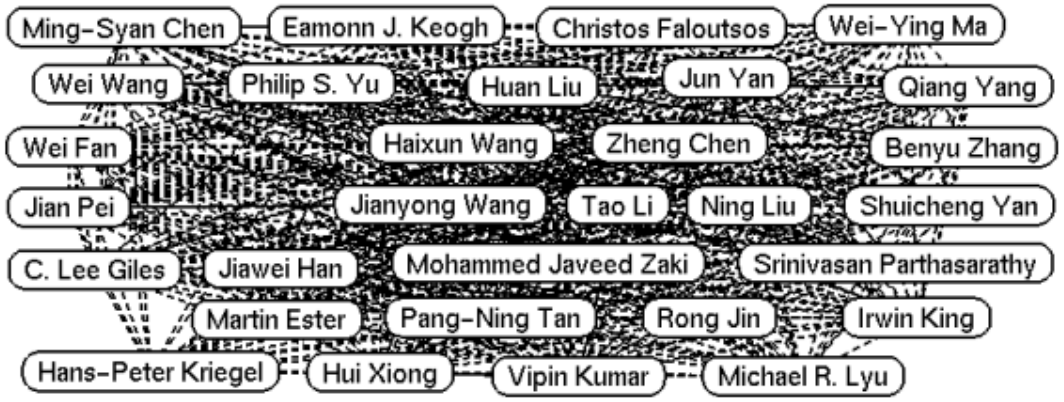
TABLE II
STATISTICS OF THE DUAL NETWORKS

Dual networks	Abbr.	#nodes	#edges in G_a	#edges in G_b
Research-DM	DM	7,169	14,526	30,000
Research-DB	DB	6,131	17,940	30,000
Recom-Epinions	EP	49,288	487,002	313,432
Recom-Flixster	FX	786,936	7,058,819	2,713,671
Protein-Genetic	Bio	8,468	25,715	67,744

Densest subgraph in Conceptual Network

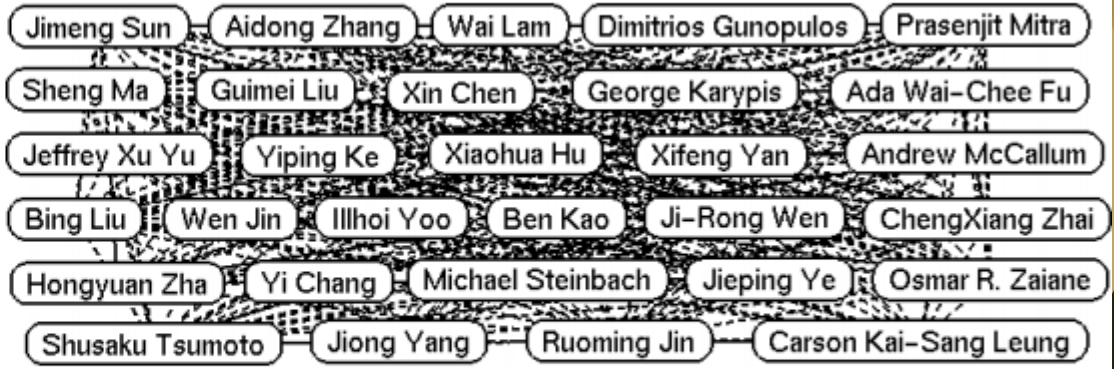


(a) Subgraph in co-author network



(b) Subgraph in research interest similarity network

(a) Induced subgraph in co-author network



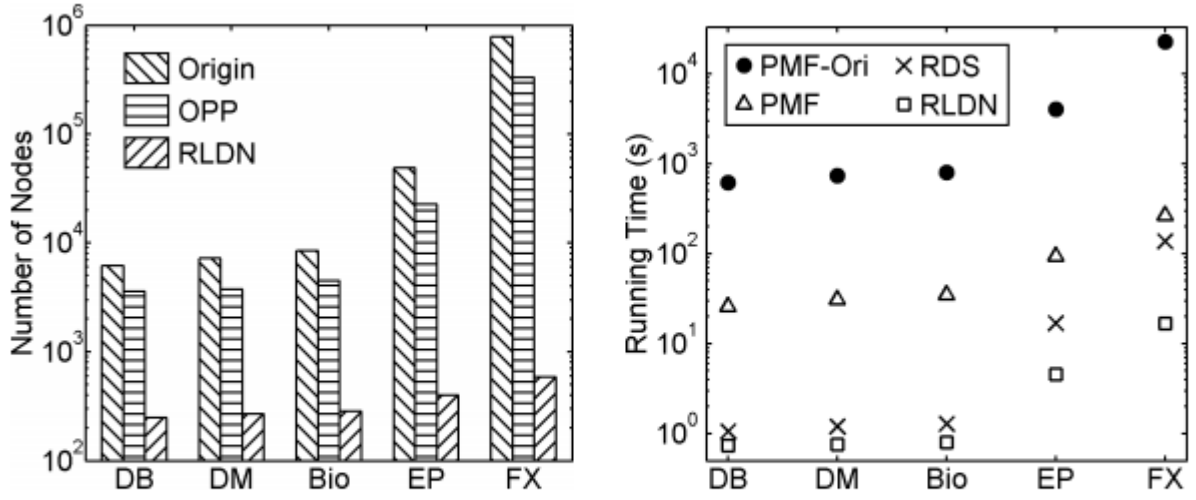
(b) Dense subgraph in research interest similarity network

Fig. 10. The dense subgraph in the research interest similarity network of the dual co-author (data mining) networks

DCS RDS

- The DCS RDS algorithm has three major components:

- Removing low degree nodes (RLDN) in the conceptual network
- Finding the densest subgraph in the remaining graph by parametric maximum flow (PMF)
- Refining the densest subgraph (RDS) to make it connected in the physical network



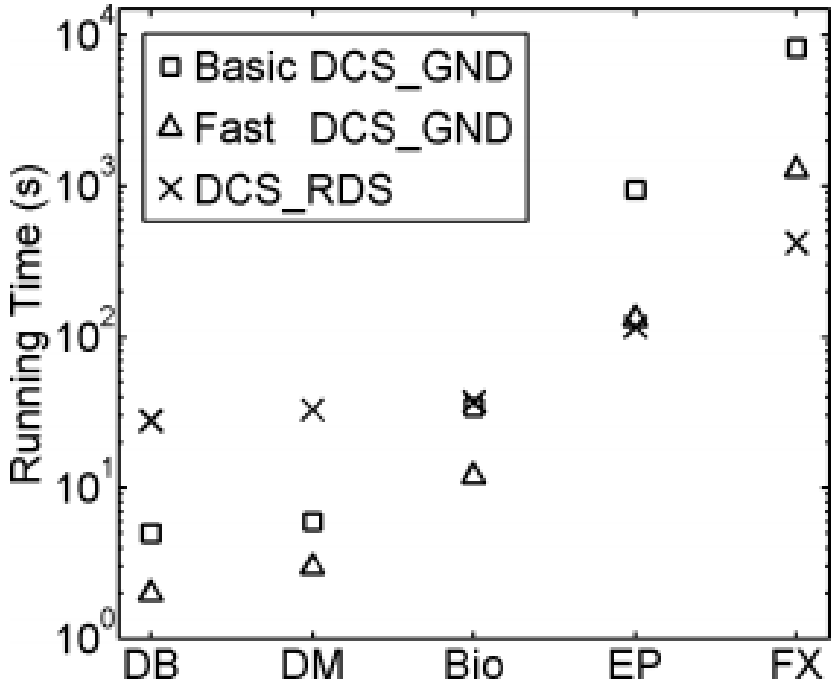
(a) Pruning effect (b) Running time

Fig. 16. Pruning effect and running time of the DCS_RDS algorithm

DCS GND Methods

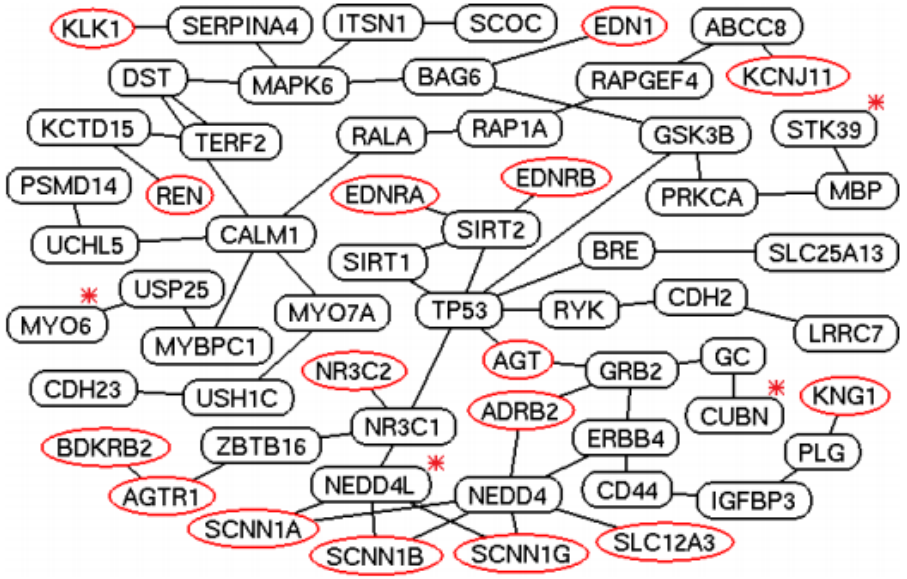
- This demonstrates the effectiveness of simultaneously deleting independent non-articulation nodes

- We can observe that DCS GND runs faster on smaller graphs and DCS RDS runs faster on larger graphs

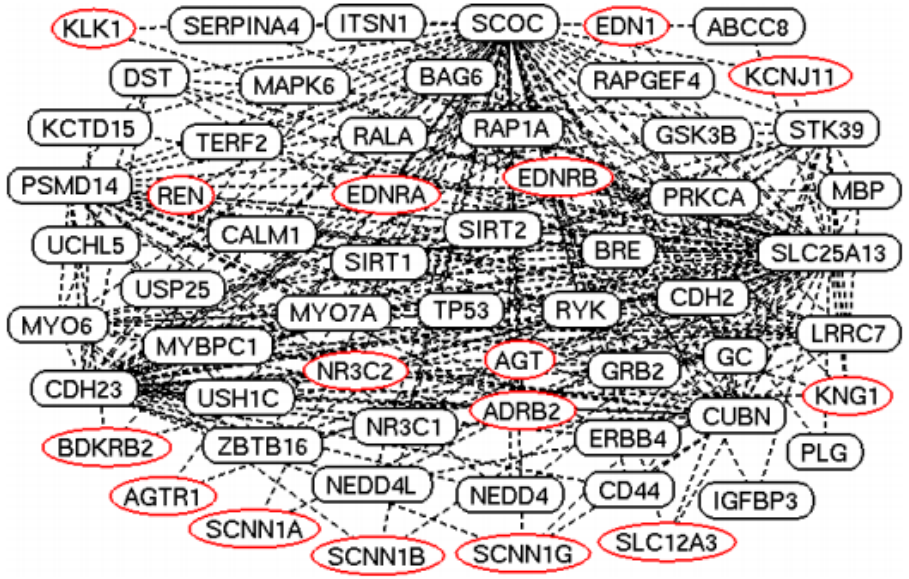


(a) Running time

Result in Genetics



(a) Subgraph in protein interaction network



(b) Subgraph in genetic interaction network

Thank you…

Appendix

Lemma 1: The DCS in dual networks does not contain any low degree leaf node.

Proof: Suppose otherwise. We remove u from S and let S' be the remaining set of nodes. Since $G_a[S]$ is connected and u is a leaf node in G_a, so after deleting u, $G_a[S']$ is still connected. However, its density $\rho(S') = \frac{|E_b(S')|}{|S'|} = \frac{|E_b(S)|-w_{G_b}(u)}{|S|-1} > \frac{|E_b(S)|}{|S|} = \rho(S)$, since $w_{G_b}(u) < \rho(S) = \frac{|E_b(S)|}{|S|}$. This contradicts the assumption. ■

Proof

- *Lemma 3:* Let $\alpha = \rho(T)/d$. The d-core subgraph D is a 2α-approximation of the densest subgraph $G[T]$.

  Proof: Let $D.V$ represent the node set in D. Since the density of the d-core is $\rho(D) = \frac{|E(D.V)|}{|D.V|} = \frac{\sum_{u \in D.V} w_D(u)}{2|D.V|} \geq \frac{\sum_{u \in D.V} d}{2|D.V|} = \frac{d}{2} = \frac{\rho(T)}{2\alpha}$, we have that $\rho(T) \leq 2\alpha\rho(D)$. ■