Identifying Protein Complexes from Protein Interactome Maps

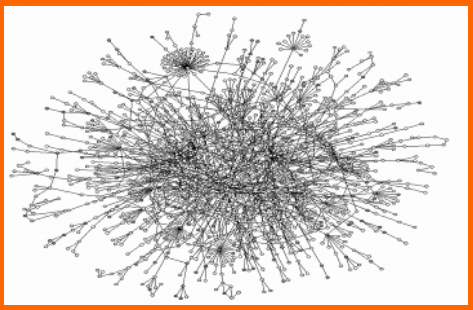
Limsoon Wong

(Joint work with Kenny Chua & Guimei Liu)

Plan

- **Motivation and Approaches**
- **PPI Network Cleansing based on PPI Topology**
  - CD-Distance, FS-Weight
- **Impact of Cleansing on PPI-based Protein Complex Prediction Methods**
- **Recent Improvement to PPI Network Cleansing and PPI-Based Protein Complex Prediction**

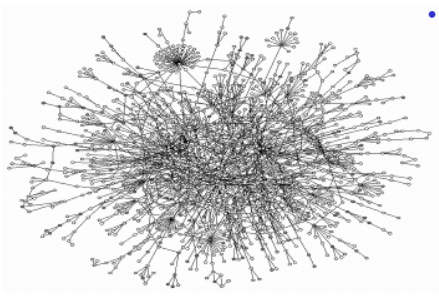
Motivation & Approach



NUS National University of Singapore

Motivation

- **Nature of high-throughput PPI expts**
  - Proteins are taken out of their natural context!



- **Can a protein interact with so many proteins simultaneously?**
- **A big "hub" and its "spokes" should probably be decomposed into subclusters**
  - Each subcluster is a set proteins that interact in the same space and time
  - Viz., a protein complex

Approaches to PPI-Based Protein Complex Prediction

	RNSC	MCODE	MCL
Type	Clustering, local search cost based	Local neighborhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

- Issue: recall vs precision has to be improved

Possible Cause of Low Recall/Precision

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

Sprinzak et al., *JMB*, 327:919-923, 2003

Large disagreement betw methods

- High level of noise

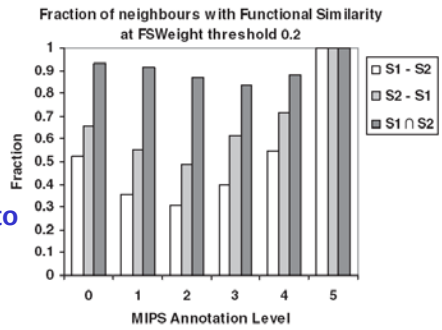
⇒ Need to clean up before protein complex prediction

PPI Network Cleansing based on PPI Topology

Guilt by Association of Common Interaction Partners

- **Two proteins participating in same biological process/ cellular compartment are likely to interact**
- **Two proteins having a large proportion of their interaction partners in common are likely to participate in same biological process/ cellular compartment**

⇒ **Two proteins having a large proportion of their interaction partners in common are likely to directly interact also**



Chua et al, Bioinformatics, 22:1623--1630, July 2006

Measures that correlate with function homogeneity and localization coherence

- **Two proteins participating in same biological process are more likely to interact**
- **Two proteins in the same cellular compartments are more likely to interact**

- **CD-distance & FS-Weight**
  - Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment
  - Correlate well with functional homogeneity and localization coherence

Czekanowski-Dice Distance

- **Functional distance between two proteins** (Brun et al, 2003)

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$

- **N_k is the set of interacting partners of k**
- **X Δ Y is symmetric diff betw two sets X and Y**
- **Greater weight given to similarity**

⇒ **Similarity can be defined as**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y + Z)}$$

FS-Weight

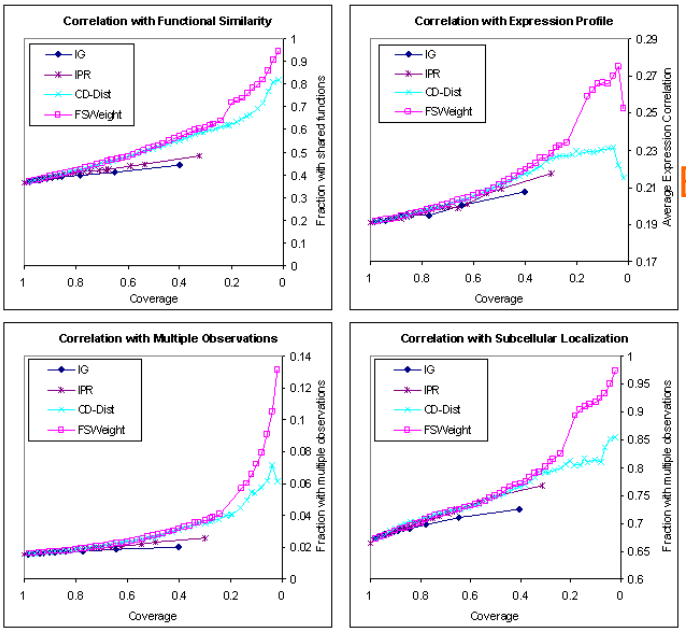
- **FS-Weight** (Chua et al, 2006)

$$S(u,v)=\frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v|} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v|}$$

  - N_k is the set of interacting partners of k
  - Greater weight given to similarity

⇒ **Rewriting this as**

$$S(u,v)=\frac{2X}{2X+Y} \times \frac{2X}{2X+Z}$$



Evaluation wrt Common Cellular Role,etc

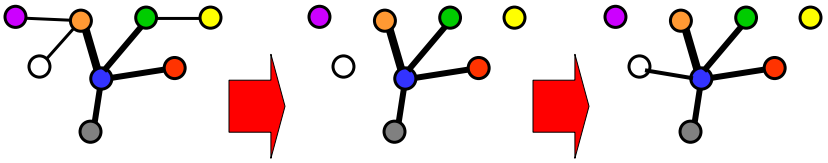
Impact of Cleansing on PPI-Based Protein Complex Prediction Methods

PPI-Based Complex Prediction Algo

	RNSC	MCODE	MCL
Type	Clustering, local search cost based	Local neighborhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

- Issue: recall vs precision has to be improved
- Does a "cleaner" PPI network help?

Cleaning PPI Network by FS-Weight



- **Modify existing PPI network as follow**
  - Remove level-1 interactions with low FS-weight
  - Add level-2 interactions with high FS-weight
- **Then run RNSC, MCODE, MCL, & PCP**

Experiments

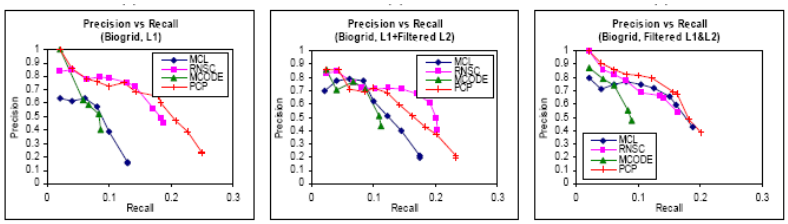
- **PPI datasets**
  - PPI[BioGRID], BioGRID db from Stark et al., 2006
- **Gold standards**
  - PC_{2004}, Protein complexes from MIPS 03/30/2004
  - PC_{2006}, Protein complexes from MIPS 05/18/2006
- **Validation criteria**

$$overlap(S,C) = \frac{|V_S \cap V_C|^2}{|V_S| \cdot |V_C|}$$

where

  - S = predicted cluster
  - C = true complex
  - V_x = vertices of subgraph defined by X
- **Overlap(S,C) ≥ 0.25 is considered a correct prediction**

Validation on PC_{2004}



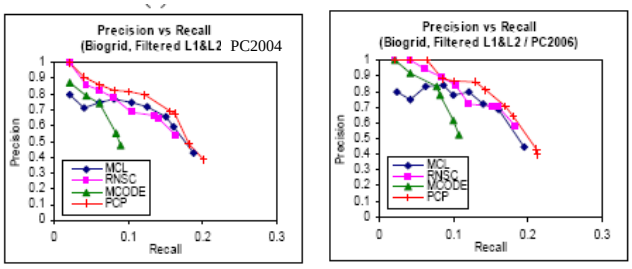
d) Original level-1 PPI

e) Original level-1 PPI and filtered level-2 PPI

f) Filtered level-1 and level-2 PPI

- **Precision is improved in all methods**

Validation on PC_{2006}



- **When predictions are validated against PC_{2006}, precision of all algo improved**
- **Many "false positives" wrt PC_{2004} are actually real**

PCP Algorithm

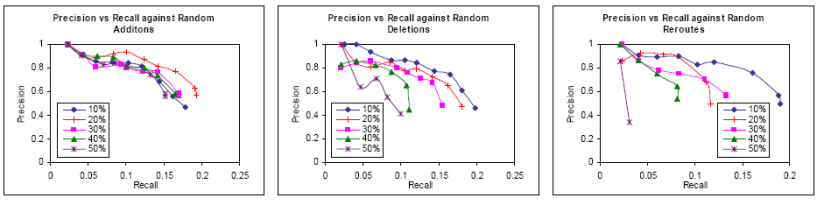
- **Find all max cliques in the modified PPI network**
  - If two cliques overlap, distribute the overlapped nodes such that both cliques have larger average FS-weight
- **Merge resulting (partial) cliques with good inter-cluster density**

$$ICD(S_a, S_b) = \frac{\sum S_{FS}(i,j) \mid i \in (V_a - V_b), j \in (V_b - V_a), (i,j) \in E}{|V_a - V_b| \cdot |V_b - V_a|}$$

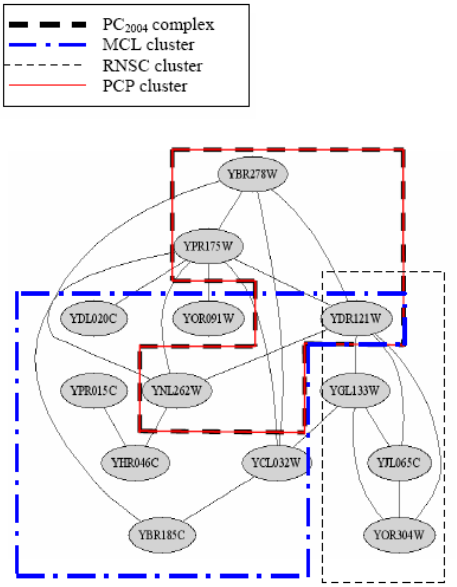
- **Modify the PPI network by treating the merged partial cliques as vertices**
- **Iterate the steps above**

Chua et al, *JBCB*, 6:435-466, 2008

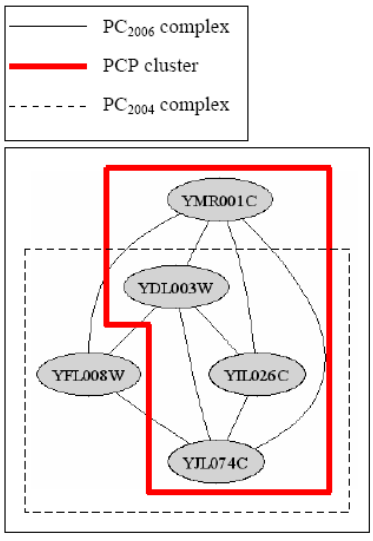
Robustness of PCP Against Noise



- **PCP is robust against 10-50% random additions**
  - FW-weight is able to remove spurious interactions
- **Random deletions negatively impacts recall**
  - Increased sparseness caused edges to received smaller FS-weight; more interactions got filtered
  - Led to insufficient info to form good cliques



PCP Prediction Example 1



PCP Prediction Example 2

Conclusions

- **Precision of protein complex prediction can be improved by**
  - PPI network augmented with level-2 interactions
  - PPI network cleansed by FS-weight

Recent Improvement to PPI Network Cleansing & PPI-Based Protein Complex Prediction

Expectation Maximization

- **CD-distance**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y + Z)}$$

- **X is # common neighbours of 1st & 2nd proteins**
- **Y/Z is # unique neighbours of 1st/2nd protein**

- **These counts are noisy**
- **Use CD-distance to weigh these counts and recompute CD-distance**

⇒ **Iterated CD-distance, ditto for FS-weight**

Local Score: Iterated CD-Distance

- **CD-Distance**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y + Z)}$$

- **Define "local score" by iterating CD-distance**

$$w_L^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} w_L^{k-1}(x,u) + \sum_{x \in N_u \cap N_v} w_L^{k-1}(x,v)}{\sum_{x \in N_u} w_L^{k-1}(x,u) + \sum_{x \in N_v} w_L^{k-1}(x,v) + \lambda_u^k + \lambda_v^k}$$

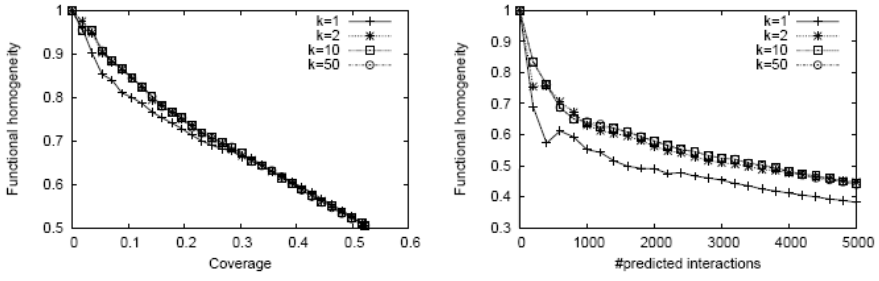
where $w_L^{k-1}(x,u)$ is the score of (x,u) in the $(k\text{-}1)$-th iteration, $w_L^0(x,u)=1$ if $(x,u) \in E$ and $w_L^0(x,u)=0$ if $(x,u) \notin E$. The two terms, λ_u^k and λ_v^k, are used to penalize proteins with very few neighbors

Validation of Iterated CD-Distance

- **DIP yeast dataset**
  - Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  - Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs
- **Let's see how much better iterated CD-distance/ FS-weight is over the baseline above, as well as over the original CD-distance/ FS-weight**

Performance of Iterated CD-Distance wrt Functional Homogeneity

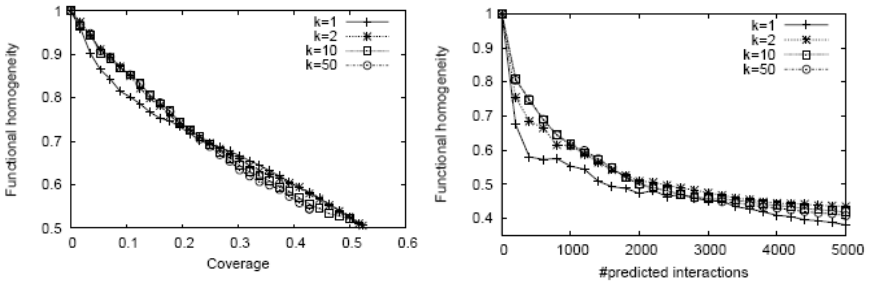
Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%



- **Iterated CD-distance achieves best performance wrt functional homogeneity at k=2**
- **Ditto wrt localization coherence (not shown)**

Performance of Iterated FS-Weight wrt Functional Homogeneity

Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%

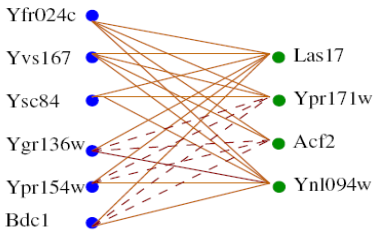


- **Iterated FS-weight achieves best performance wrt functional homogeneity at k=2**
- **Ditto wrt localization coherence (not shown)**

Interacting Motif Pairs

- **If one group of proteins interact with another group, it is likely that the interactions are mediated by an underlying complementary domain/motif pair**

⇒ **If a protein pair participates in an interacting protein group pair whose two groups are densely connected, then the interaction between these two proteins is more likely to be true**



Global Score

- **Find all protein group pairs V_1 and V_2 that are sufficiently large and have dense interactions**
- **Then define "global score" of a PPI based on interacting confidence of the interacting group pairs it participates in and the degree of its participation**

$$w_G(u,v) = \max\{conf(V_1,V_2) \cdot \frac{2|N_u \cap V_2|}{|V_2|+|N_u|} \cdot \frac{2|N_v \cap V_1|}{|V_1|+|N_v|} \mid u \in V_1, v \in V_2\}$$

where conf(V_1, V_2) is the ratio of # of interactions betw V_1 and V_2 to # of distinct protein pairs contained in (V_1, V_2)

Combining Local and Global Scores

- **Local score**

$$w_L^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} w_L^{k-1}(x,u) + \sum_{x \in N_u \cap N_v} w_L^{k-1}(x,v)}{\sum_{x \in N_u} w_L^{k-1}(x,u) + \sum_{x \in N_v} w_L^{k-1}(x,v) + \lambda_u^k + \lambda_v^k}$$

- **Global score**

$$w_G(u,v) = \max\{conf(V_1,V_2) \cdot \frac{2|N_u \cap V_2|}{|V_2|+|N_u|} \cdot \frac{2|N_v \cap V_1|}{|V_1|+|N_v|} \mid u \in V_1, v \in V_2\}$$

Where conf(V_1, V_2) is the ratio of # of interactions betw V_1 and V_2 to # of distinct protein pairs contained in (V_1, V_2)

- **Combined measure**

$$LGTweight(u,v) = w_L^2(u,v) + w_G(u,v).$$

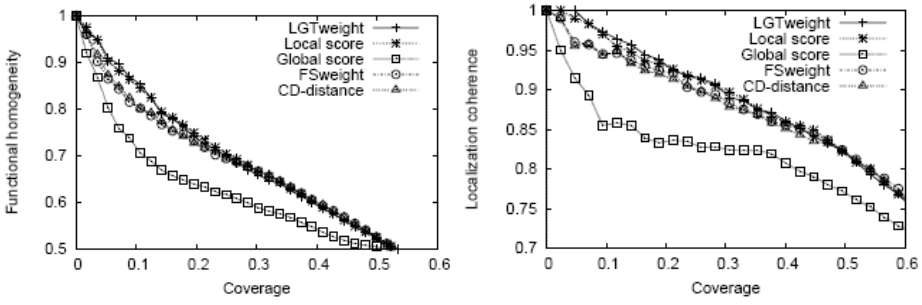
Liu & Wong, GIW 2008

Validating the Scores

- **Retain protein group pairs (V_1, V_2) where**
  - V_1, V_2 have at least 5 members each
  - V_1 proteins have ≥ 1 common partner in V_2
  - Conf(V_1,V_2) ≥ 0.1
  - P-value for such protein group pairs is < 0.005
- **Use DIP yeast dataset to check functional homogeneity and localization coherence of PPI ranked by our scores**
- **Use DIP yeast dataset in 5-fold cross validation**

Identifying False Positive PPIs

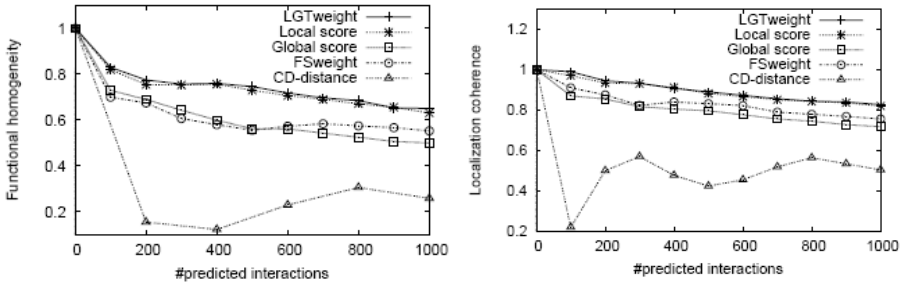
Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



- **LGTweight & iterated CD-distance are improvement over previous measures for assessing PPI reliability**

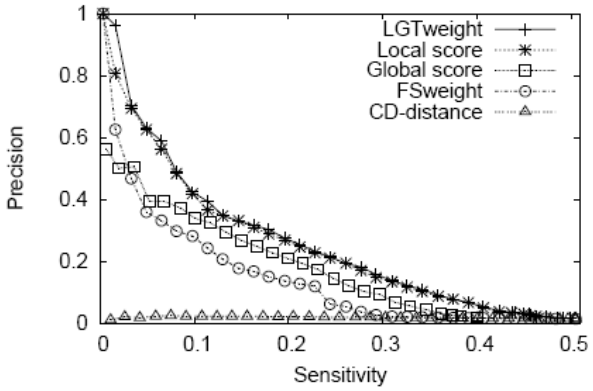
Identifying False Negative PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



- **LGTweight & iterated CD-distance are improvement over previous measures for predicting new PPIs**

5-Fold X-Validation



- **LGTweight, iterated CD-distance, & our global score are improvement over previous measures for identifying false positive & false negative PPIs**

Now we can make protein complex prediction as follows…

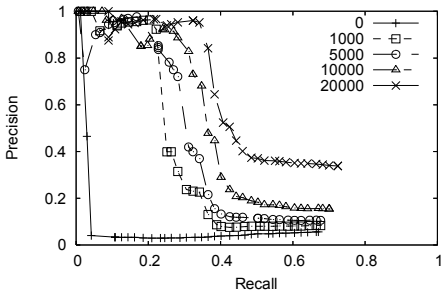
- **Remove noise edges in the input PPI network by discarding edges having low score/LGTweight**
- **Augment the input PPI network by addition of missing edges having high score/LGTweight**
- **Predict protein complex by simple clique finding!**

Validation Experiments

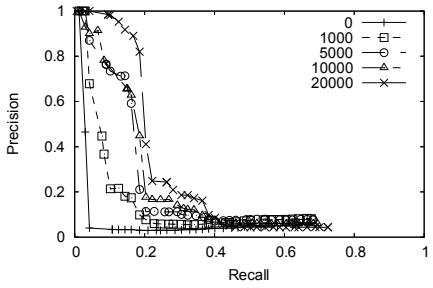
- **Matching a predicted complex S with a true complex C**
  - Vs: set of proteins in S
  - Vc: set of proteins in C
  - Overlap(S, C) = $|Vs \cap Vc|^2 / |Vs||Vc|$
  - Overlap(S, C) ≥ 0.25
- **Evaluation**
  - Precision = matched predictions / total predictions
  - Recall = matched complexes / total complexes
- **Datasets: BioGrid yeast**
  - #interactions: 38555
  - #interactions with >0 common neighbor: 27940

FSweight vs Iterated FSweight

- **Remove interactions that have no common neighbors**
- **Add N new interactions to the remaining network, N=0, 1000, 5000, 10000, 20000**



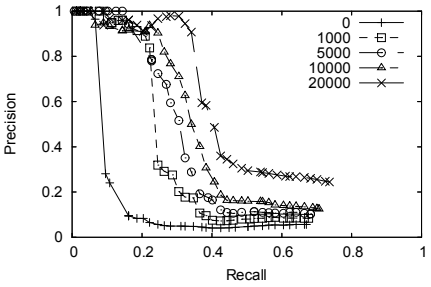
Iterated FSweight



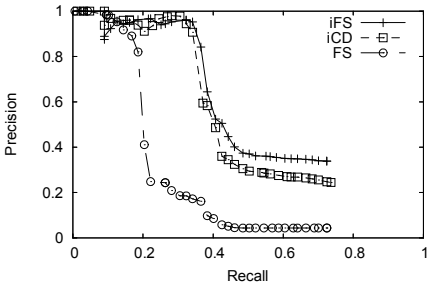
FSweight

Iterated CD-Distance vs FSWeight

- **Remove interactions that have no common neighbors**
- **Add N new interactions to the remaining network, N=0, 1000, 5000, 10000, 20000**



Iterative adjusted CD-distance



20000 new interactions added

What have we learned?

- **Guilt by association of common interaction partners is useful for predicting**
  - PPI cellular localization
  - Missing PPIs
  - Protein complexes

- **Acknowledgement**
  - Kenny Chua, Guimei Liu

Readings

- **H.N. Chua, et al. Using Indirect Protein-Protein Interactions for Protein Complex Prediction. *Journal of Bioinformatics and Computational Biology*, 6(3):435--466, 2008**

- **G. Liu, J. Li, L. Wong. "Assessing and predicting protein interactions using both local and global network topological metrics", *Proc GIW2008***

Any Question?