Link Prediction in Social Networks

CS5128701:
Practice of Social Media Analytics

NTUST CSIE
Yi-Ling Chen

What is link prediction?

- **Link prediction**. A network is changing over time. Given a snapshot of a network at time t, predict edges added in the interval (t, t')

Link prediction by proximity scoring

1. For each pair of nodes compute proximity (similarity) score $c(v_1, v_2)$
2. Sort all pairs by the decreasing score
3. Select top n pairs (or above some threshold) as new links

threshold
0.6

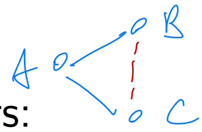
$V_2\ V_7\ \ 0.9$
$V_3\ V_8\ \ 0.7$
$V_6\ V_{12}\ \ 0.5$
$V_9\ V_3\ \ 0.3$

(Slide Credit: Leonid Zhukov)

Scoring Function

transitivity

Local neighborhood of v_i and v_j



越多 common neighbors
BC 越容易認識

- Number of common neighbors:

$$\overset{A}{|\mathcal{N}(v_i)} \cap \overset{B}{\mathcal{N}(v_j)|}$$

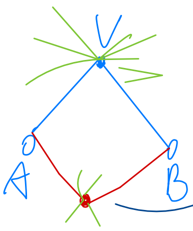
- Jaccard's coefficient:

$$\frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)|}$$

N_1 : 10
AB common neighbors 5 $\frac{5}{10}$
N_2 : 1000
AB common neighbors 10 $\frac{10}{1000}$

- Adamic/Adar: 針對不同 common friends 給不同 weight

$$\sum_{v \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{1}{\log |\mathcal{N}(v)|}$$



給它較 high weight

越 low 越好
$\log|N(v)| \downarrow \Rightarrow N(v) \downarrow$.

(Slide Credit: Leonid Zhukov)

Scoring Function (cont'd)

Paths and ensembles of paths between v_i and v_j

- Shortest path:

$$-\min_{s}\{path_{ij}^{s} > 0\}$$

7 10 15 ⇒ -7
∴ -7 > -30 所以要加負号

- Katz score:

$0 < \beta < 1$

$$\sum_{l=1}^{\infty} \beta^l |paths^{(l)}(v_i, v_j)|$$

$\beta = 0.5$

7條: 2, 3, 3, 5, 5, 5, 17

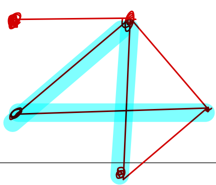
$\sum_{l=2,3,5,17} 0.5^2 \times 1 + 0.5^3 \times 2 + 0.5^5 \times 3 + 0.5^{17} \times 1$

較小

- Expected number of random walk steps:
  - hitting time: $-H_{ij}$

V_i 8 / 4 / 10 V_j

$\frac{8+4+10}{3}$



(Slide Credit: Leonid Zhukov)

Scoring Function (cont'd)

- Preferential attachment:

$$k_i \cdot k_j = |\mathcal{N}(v_i)| \cdot |\mathcal{N}(v_j)|$$

or

$$k_i + k_j = |\mathcal{N}(v_i)| + |\mathcal{N}(v_j)|$$

- Clustering coefficient:

$$CC(v_i) \cdot CC(v_j)$$

or

$$CC(v_i) + CC(v_j)$$

(Slide Credit: Leonid Zhukov)

Efficient Prediction and Recommendation in Social Networks

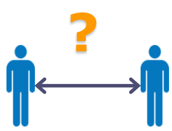
Recommendations in Social Networks

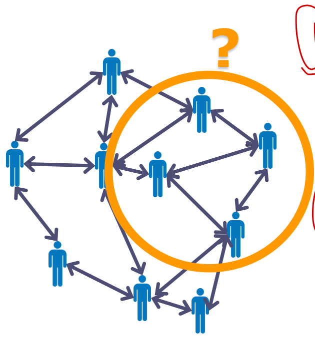
- Social networks are very informative, and we can provide users many helpful suggestions with social networks.
  - social filtering [1] 把别人喜欢的也推给你
  - recommending "people you may know" (link prediction)
- However, sophisticated recommendations are often computationally expensive.

[1] Y.-M. Li, H.-W. Hsiao, and Y.-L. Lee, "Recommending social network applications via social filtering mechanisms", Information Sciences, 2013.

Recommendations in Social Networks (cont'd)

- We focus on how to efficiently provide the following two kinds of recommendations.
  - Part 1: Diverse Ensemble with Drastic Sparsification (DEDS) for Link Prediction
  - Part 2: Social-Temporal Group Query (STGQ) for Activity Planning





① social

② temporal

Part 1: Diverse Ensemble with Drastic Sparsification (DEDS)

Motivations

- Link prediction is an important application in many fields, and lots of major problems involving networks can benefit from the process.
- Examples
  - friend recommendations on social networking sites
  - suggesting “you may also like …” items to customers

Motivations (cont'd)

- However, as networks grow larger and larger nowadays, link prediction becomes more time consuming.
- If the prediction takes hours or even days to complete, the recommendations cannot be made to users in a timely manner and may therefore become less useful as time passes.

Motivations (cont'd) 稀疏化

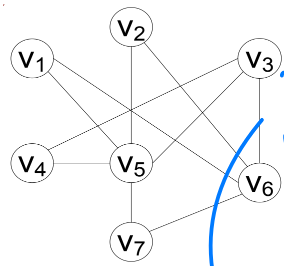
- If the network is sparsified first, the prediction cost can be considerably lowered, but the information used in link prediction also becomes scarce.

相对 information 变小

- How can we strike a good balance between efficiency and accuracy?

Dynamic network need snapshot

t_1:



feature value

	CN	JC	PA	...
(V_1,V_2)	2	1	4	...
(V_1,V_3)	2	2/3	6	...
(V_1,V_4)	1	1/3	4	...
(V_1,V_7)	2	1	4	...
...	...	...	...	...

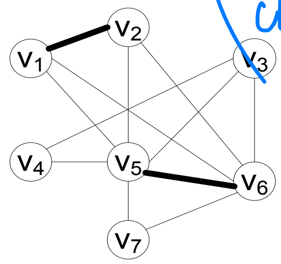
No sparsified

training data

⇓

model

t_2:



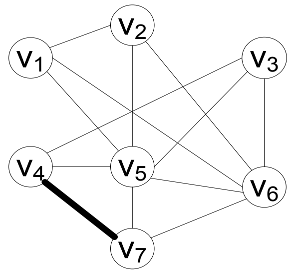
classification label

	Label
(V_1,V_2)	positive
(V_1,V_3)	negative
(V_1,V_4)	negative
...	...

M

testing data

t_3:



	CN	JC	PA	...	Prediction
(V_1,V_3)	2	1/2	9	...	yes
(V_1,V_4)	1	1/5	9	...	no
(V_1,V_7)	2	1/2	9	...	no
(V_2,V_3)	2	1/2	9	...	yes
...	...	...	...	...	...

t_1:

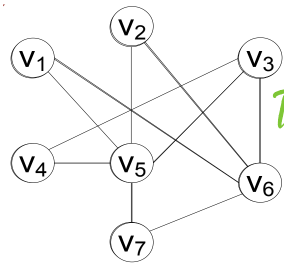


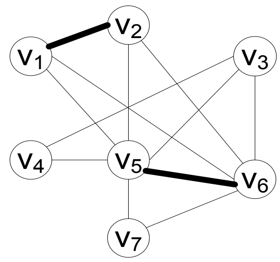
Table1

变简单

	CN	JC	PA	...
(V_1,V_2)	1	1	1	...
(V_1,V_3)	1	1/2	2	...
(V_1,V_4)	0	0	1	...
(V_1,V_7)	0	0	1	...
...	...	...	...	...

sparsification 拿掉 edges

t_2:



	Label
(V_1,V_2)	positive
(V_1,V_3)	negative
(V_1,V_4)	negative
...	...

No sparsification

Label bc value important

t_3:

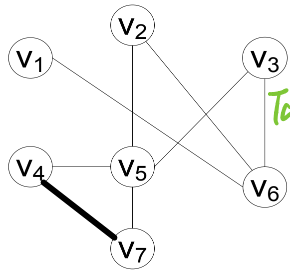


Table3

	CN	JC	PA	...	Prediction
(V_1,V_3)	1	1/2	2	...	yes
(V_1,V_4)	0	0	2	...	no
(V_1,V_7)	0	0	2	...	no
(V_2,V_3)	1	1/2	2	...	yes
...	...	...	...	...	...

if Table 1 有 spar Table 3 没有

easy Prediction Yes

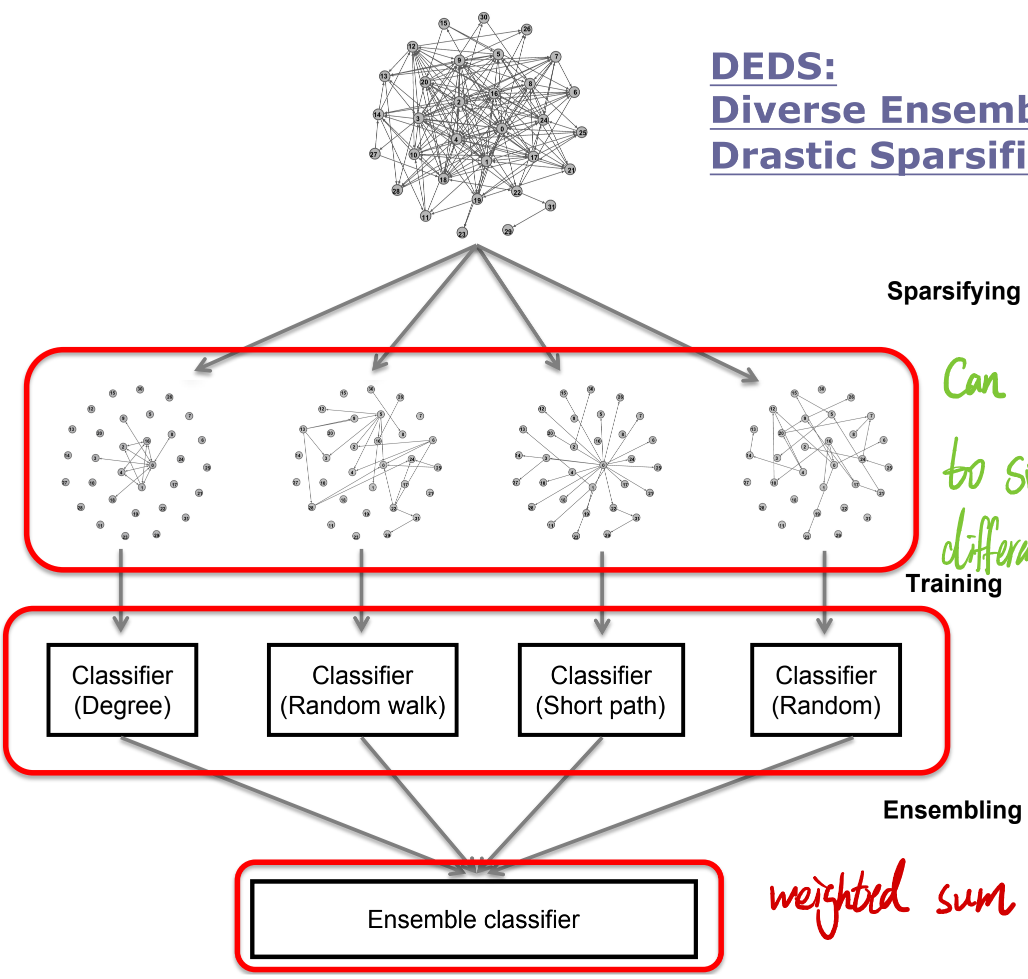
Improvement on Diversity

- According to [2], there are two important keys to obtaining a good ensemble classifier.
  - increasing diversity
  - amplifying the correct decisions 預測对可以講話大聲

C_1 — Yes
C_2 — No
C_3 — Yes
final 取 Yes

- Diversity is important in ensembling, since combining similar judgments (from similar classifiers) is not helpful for improving the final decision. 都 predict same things 有什麼用 ?

[2] R. Polikar, "Ensemble based systems in decision making", IEEE Circuits and Systems Mag., 2006.



DEDS:
Diverse Ensemble of Drastic Sparsification

Can use different to sparsifying to obtain different sparsify network

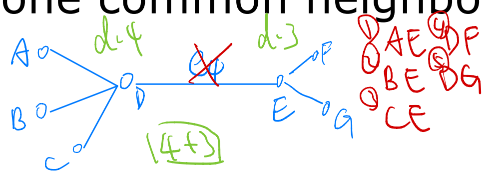
weighted sum

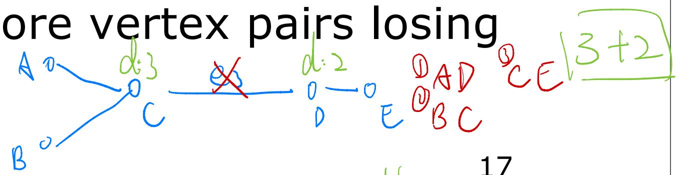
Degree-based Sparsification

- Each edge is assigned with a score in proportion to the summation of the degrees at its two ends, and the edges with higher scores are preserved.

Degree 越高越可能被保留下来

- This method is designed to reduce the loss in common neighbor counts.
  - Removing an edge with high degrees at its two ends is likely to cause more vertex pairs losing one common neighbor.





e4 will loss more common neighbors

Random-walk-based Sparsification

- Each edge is assigned with a score in proportion to the total visited count during the rehearsal random walks, and the edges with higher scores are preserved.

- This method is designed to preserve the edges most commonly used in features based on random walks between the non-neighboring vertex pair (e.g., hitting time, rooted PageRank, and PropFlow).

Short-path-based Sparsification

- Each edge is assigned with a score in proportion to the number of short paths that include this edge, and the edges with higher scores are preserved.

- This method is designed to preserve the edges most commonly used in features based on short paths between the non-neighboring vertex pair (e.g., Katz). 使 information loss 較小
  - These edges are likely to be shortcuts and important bridges. 所以可以delete

Len = 1 ~ ∞

Len = 35 $\beta^L = 0.5^{35}$ weight 會變很小

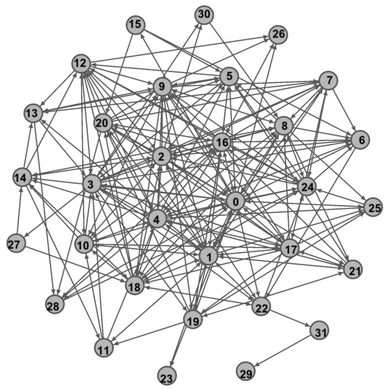
Random Sparsification

- In this sparsification method, the edges to preserve are selected randomly without any preference.

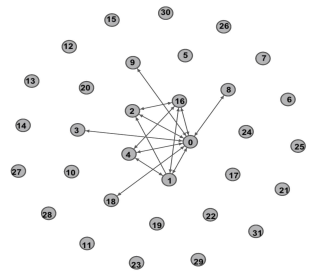
- This is the sparsification method that keeps the DEDS framework from being over-fitted to any specific feature, and also allows the DEDS framework to be more generalized for potential new features added in the future.

有各種 unique pattern

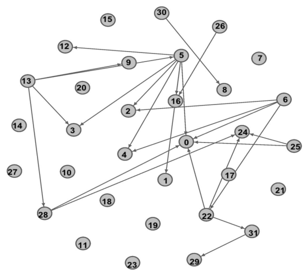
Different Sparsified Networks



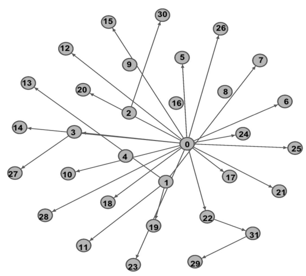
original



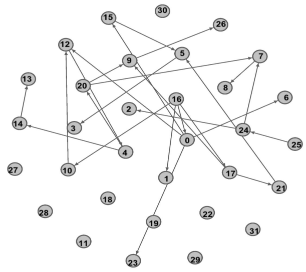
Degree



Random Walk



Short Path



Random

RSD ↑ diversity ↑

Statistics of Network Characteristics

相似性

$RSD = SD / |Mean|$

各樣

Characteristics	RSD	
	random_spar	diverse_spar
Assortativity Coef.	7%	86%
Average Clustering Coef.	17%	69%
Median Degree	0%	200%
Max Degree	4%	48%
Number of SCCs	6%	35%
Largest SCC	3%	81%
Largest SCC Diameter	17%	43%

Weights of Classifiers

- From [3], we know the additional error to the Bayes optimum decision is in proportion to the variance of the classifier, i.e., σ^2_η.
- The variance of the ensemble classifier is

要越小

$$\boxed{\sigma^2_{\eta^E} = \sum_{i=1}^{k} w_i^2 \sigma^2_{\eta^i} / \left(\sum_{i=1}^{k} w_i\right)^2}.$$

- In the dissertation, we prove that setting the weights based on the performance of classifiers (i.e., $w_i = d / \sigma^2_{\eta^i}$) can lead to better performance than equal weighting.

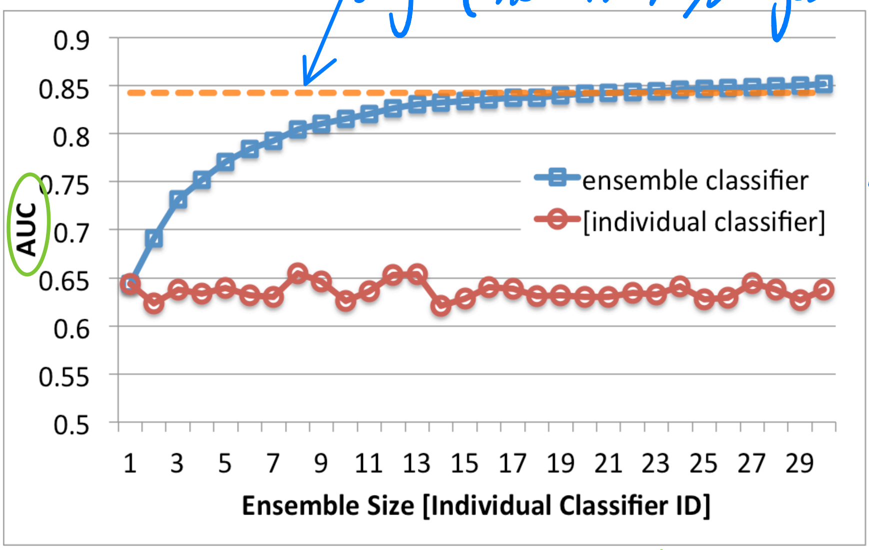
小 大

σ能就越小

[3] K. Tumer and J. Ghosh, "Theoretical foundations of linear and order statistics combiners for neural pattern classifiers", Neural Networks, 1996.

Analysis on Ensemble Size

original network 100% edges

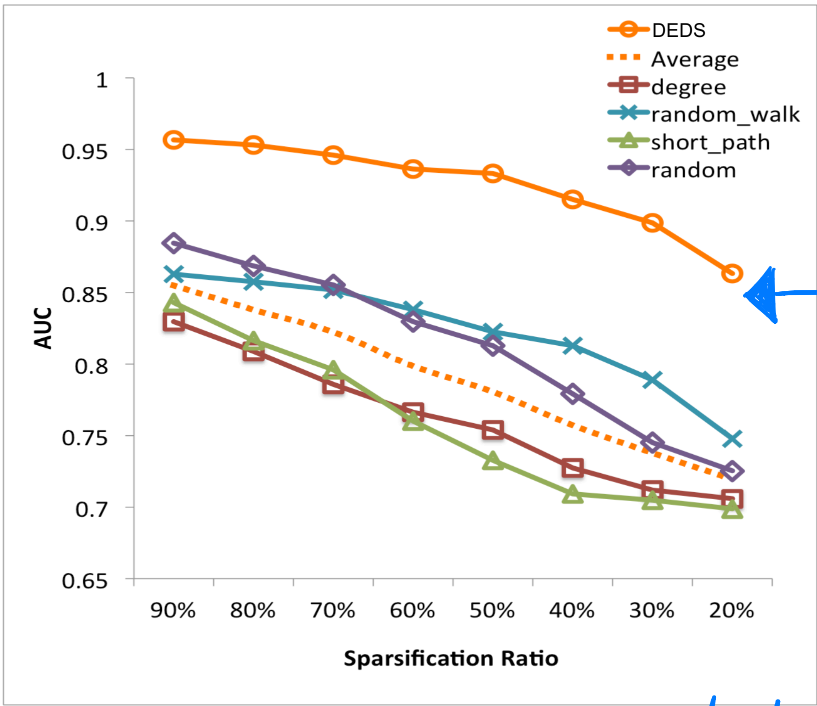


area under ROC curve

微調前四種參數產生更多 network 來 training

(sparsification ratio = 15%) if have 1000 edge will become 150

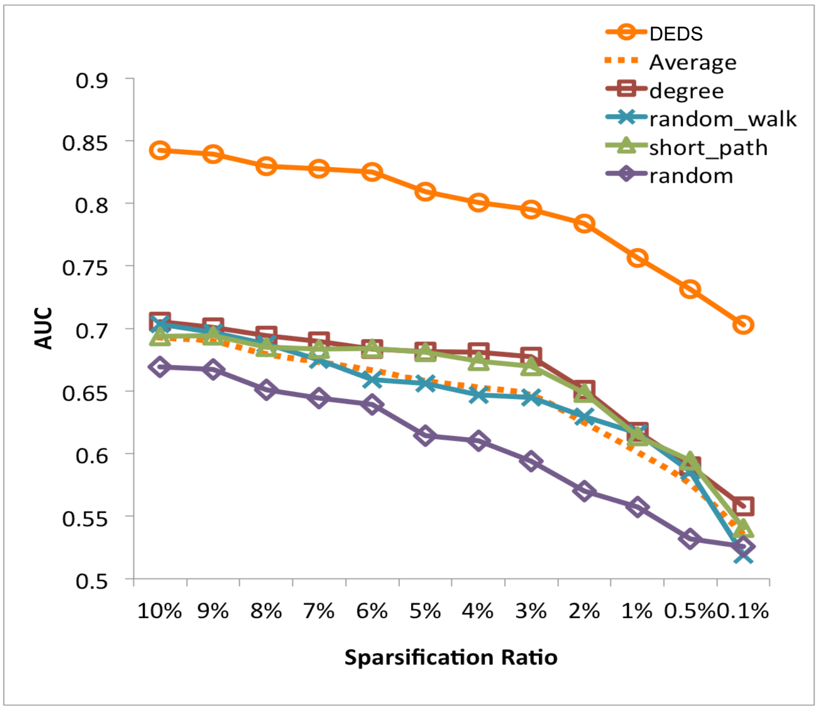
Analysis on Sparsification Ratio



用4種來做

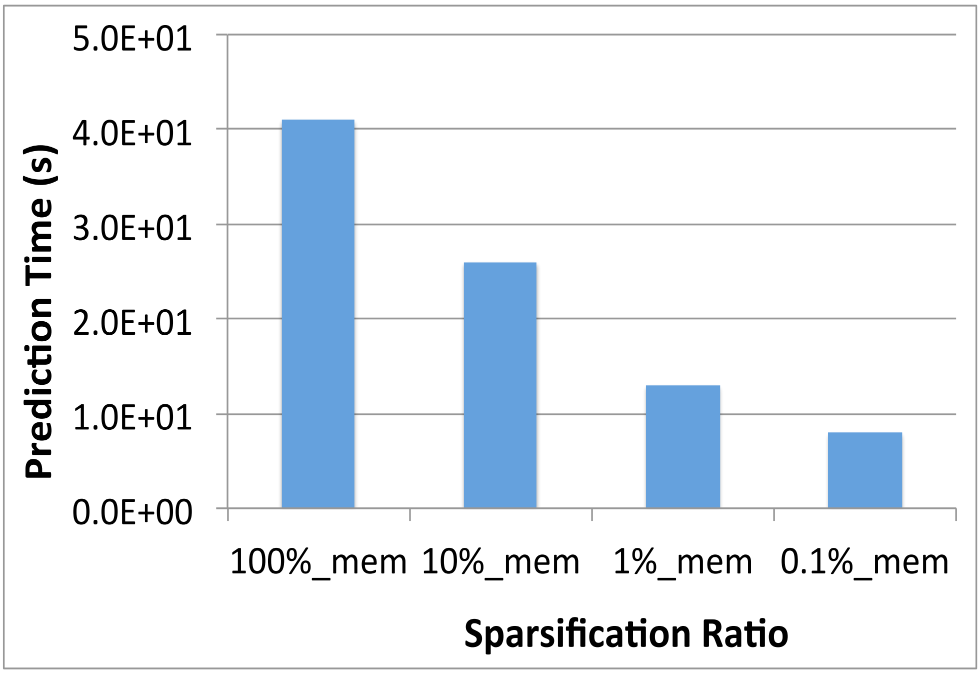
(ensemble size = 5) 5 individual

Analysis on Sparsification Ratio (cont'd)

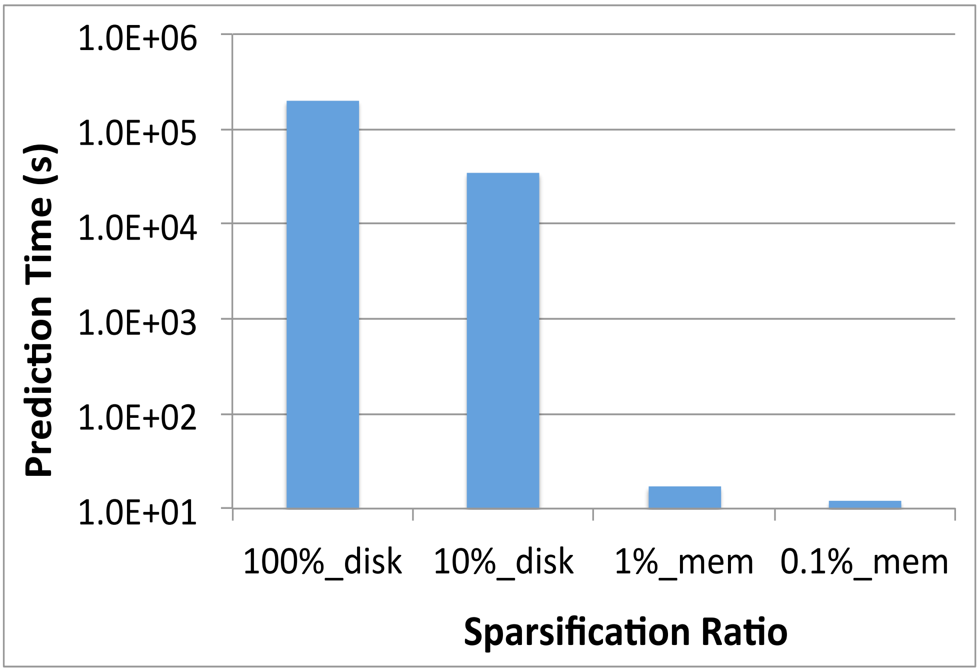


(ensemble size = 30)

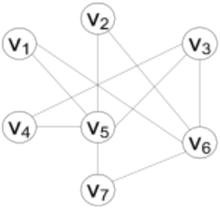
Analysis on Prediction Time



Analysis on Prediction Time (cont'd)

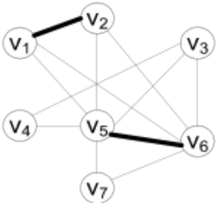


t1 :



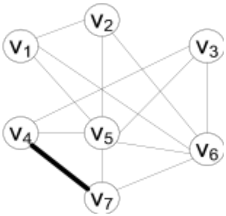
	CN	JC	PA	...
(V_1,V_2)	2	1	4	...
(V_1,V_3)	2	2/3	6	...
(V_1,V_4)	1	1/3	4	...
(V_1,V_7)	2	1	4	...
...	...	...	...	...

t2 :



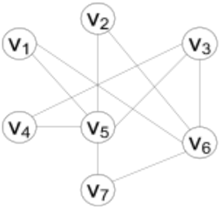
	Label
(V_1,V_2)	positive
(V_1,V_3)	negative
(V_1,V_4)	negative
...	...

t3 :



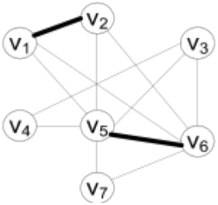
	CN	JC	PA	...	Prediction
(V_1,V_3)	2	1/2	9	...	yes
(V_1,V_4)	1	1/5	9	...	no
(V_1,V_7)	2	1/2	9	...	no
(V_2,V_3)	2	1/2	9	...	yes
...	...	...	...	...	...

t1 :



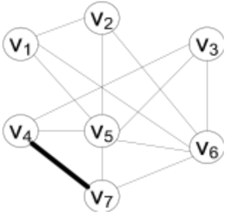
	CN	JC	PA	...
(V_1,V_2)	2	1	4	...
(V_1,V_3)	2	2/3	6	...
(V_1,V_4)	1	1/3	4	...
(V_1,V_7)	2	1	4	...
...	...	...	...	...

t2 :



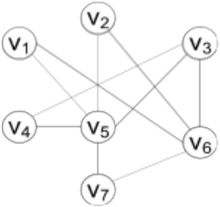
	Label
(V_1,V_2)	positive
(V_1,V_3)	negative
(V_1,V_4)	negative
...	...

t3 :



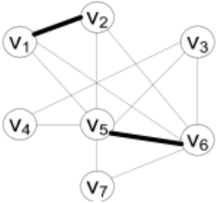
	CN	JC	PA	...	Prediction
(V_1,V_3)	2	1/2	9	...	yes
(V_1,V_4)	1	1/5	9	...	no
(V_1,V_7)	2	1/2	9	...	no
(V_2,V_3)	2	1/2	9	...	yes
...	...	...	...	...	...

t1



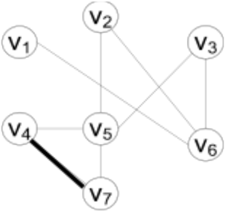
	CN	JC	PA	...
(V_1,V_2)	1	1	1	...
(V_1,V_3)	1	1/2	2	...
(V_1,V_4)	0	0	1	...
(V_1,V_7)	0	0	1	...
...	...	...	...	...

t2



	Label
(V_1,V_2)	positive
(V_1,V_3)	negative
(V_1,V_4)	negative
...	...

t3



	CN	JC	PA	...	Prediction
(V_1,V_3)	1	1/2	2	...	yes
(V_1,V_4)	0	0	2	...	no
(V_1,V_7)	0	0	2	...	no
(V_2,V_3)	1	1/2	2	...	yes
...	...	...	...	...	...