Who Is Who

——Entity Resolution

Paper #2(SVM) & Paper #6(HMRF-EM with C2+C6 constraints)

Team 1: Yuan Mei
Ruscassie, Raphael
Zhishan Wang
Yuxi Zhou(presenter)
He Zhu

Paper # 2 —— SVM

- Optimization

$$y_i(\vec{\mathbf{w}} \cdot \vec{\mathbf{x}}_i + w_0) - 1 \geq 0, \forall i$$

- Linear decision function

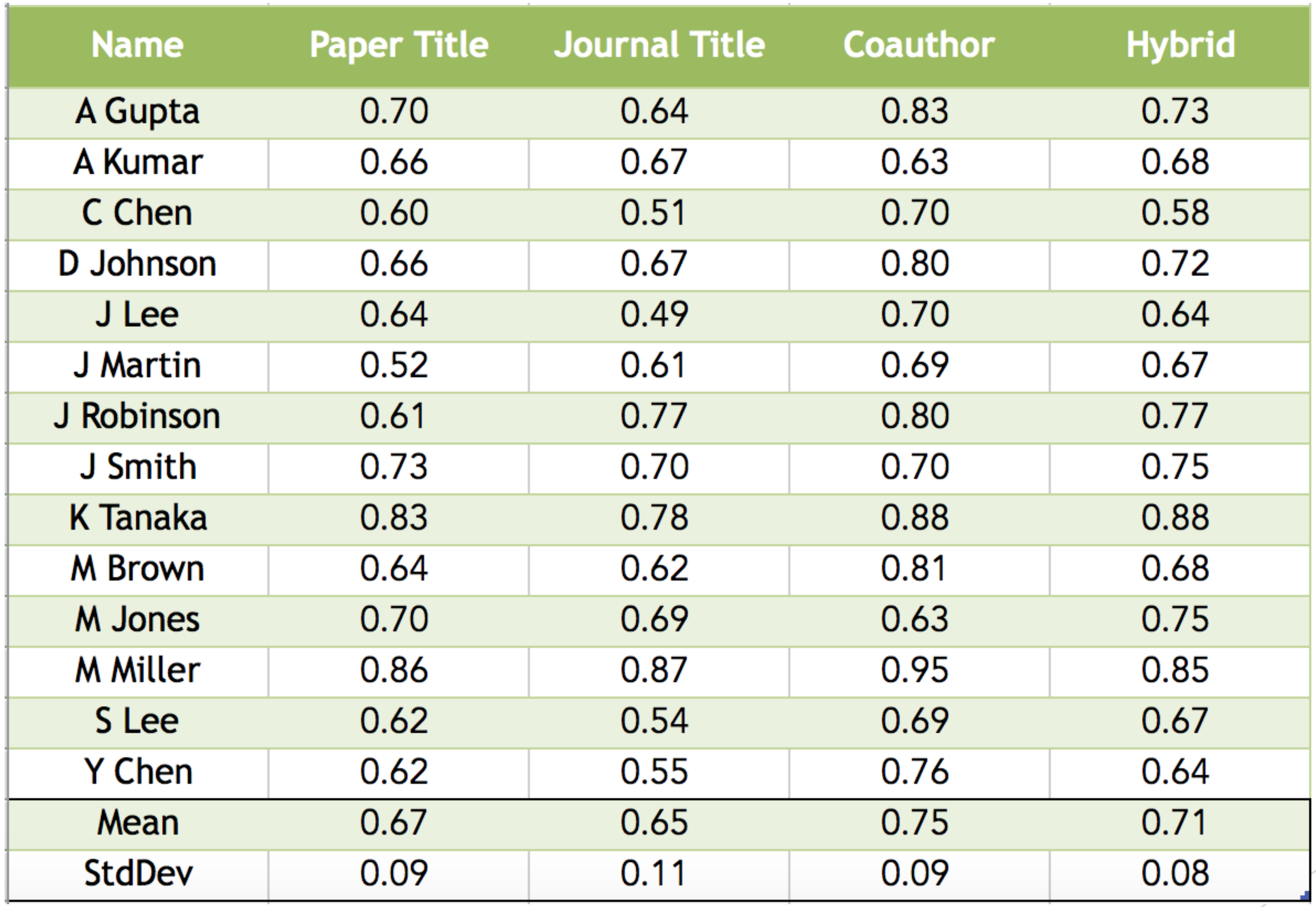
$$f(\vec{\mathbf{x}}) = sgn\{(\vec{\mathbf{w}} \cdot \vec{\mathbf{x}}) + w_0\} = sgn\{\sum_i^n \alpha_i^* y_i(\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}) + w_0^*\}$$

- Classification performance

$$\sum_i^n \alpha_i^* y_i x_{ij}$$

Paper # 2 —— SVM

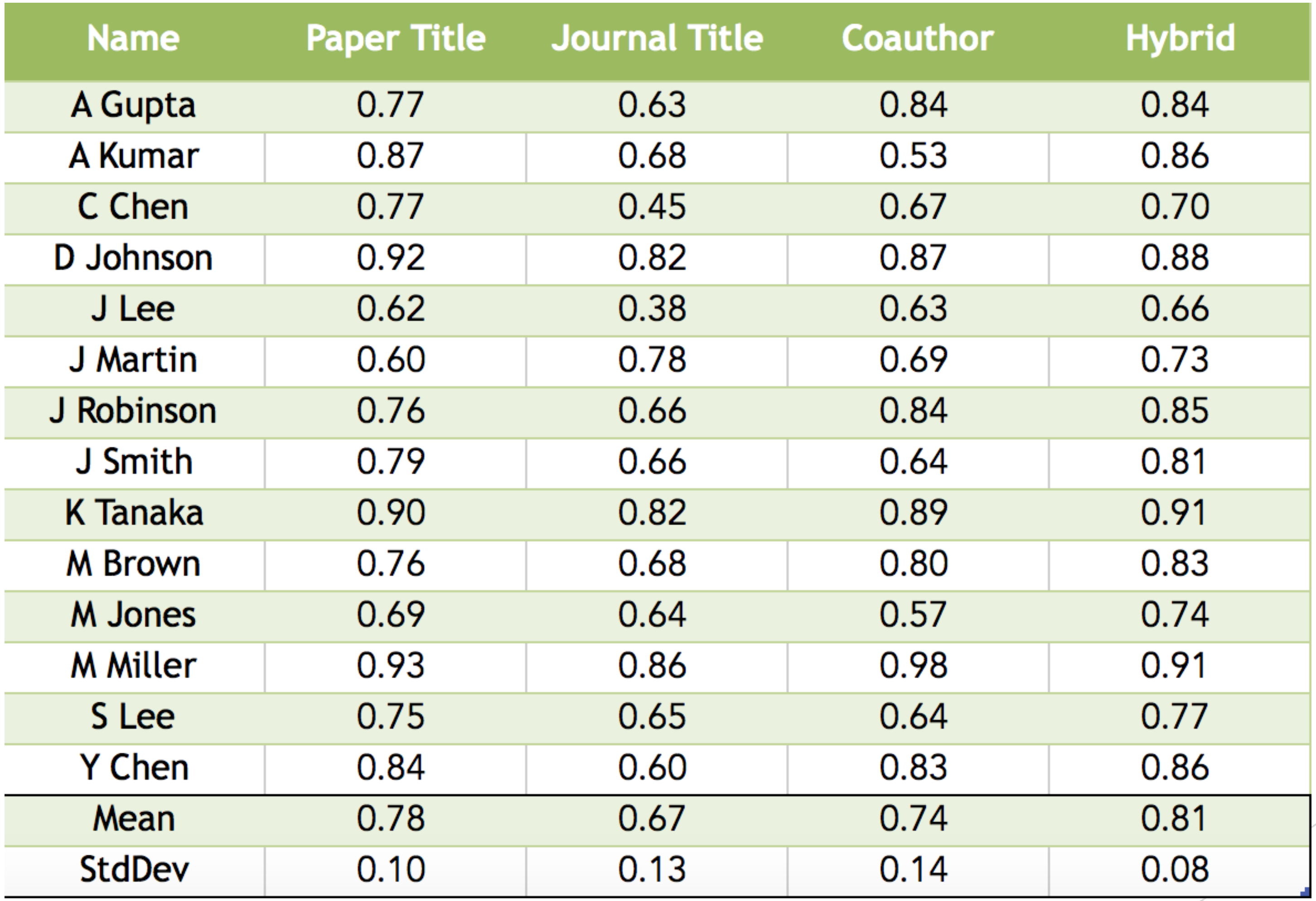
- Evaluation(accuracy)



Name	Paper Title	Journal Title	Coauthor	Hybrid
A Gupta	0.70	0.64	0.83	0.73
A Kumar	0.66	0.67	0.63	0.68
C Chen	0.60	0.51	0.70	0.58
D Johnson	0.66	0.67	0.80	0.72
J Lee	0.64	0.49	0.70	0.64
J Martin	0.52	0.61	0.69	0.67
J Robinson	0.61	0.77	0.80	0.77
J Smith	0.73	0.70	0.70	0.75
K Tanaka	0.83	0.78	0.88	0.88
M Brown	0.64	0.62	0.81	0.68
M Jones	0.70	0.69	0.63	0.75
M Miller	0.86	0.87	0.95	0.85
S Lee	0.62	0.54	0.69	0.67
Y Chen	0.62	0.55	0.76	0.64
Mean	0.67	0.65	0.75	0.71
StdDev	0.09	0.11	0.09	0.08

Paper # 2 —— SVM

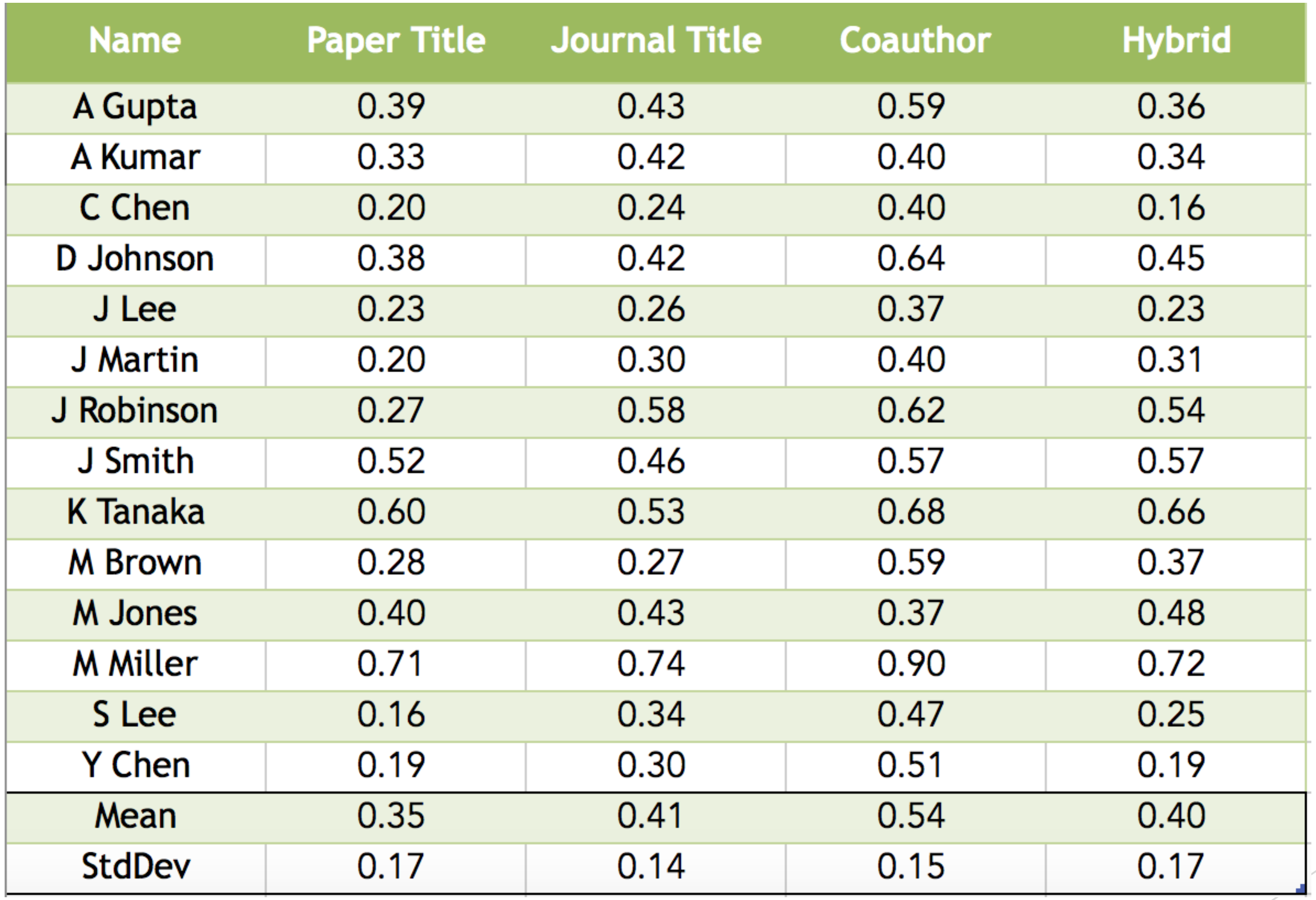
- Evaluation(precision)



Name	Paper Title	Journal Title	Coauthor	Hybrid
A Gupta	0.77	0.63	0.84	0.84
A Kumar	0.87	0.68	0.53	0.86
C Chen	0.77	0.45	0.67	0.70
D Johnson	0.92	0.82	0.87	0.88
J Lee	0.62	0.38	0.63	0.66
J Martin	0.60	0.78	0.69	0.73
J Robinson	0.76	0.66	0.84	0.85
J Smith	0.79	0.66	0.64	0.81
K Tanaka	0.90	0.82	0.89	0.91
M Brown	0.76	0.68	0.80	0.83
M Jones	0.69	0.64	0.57	0.74
M Miller	0.93	0.86	0.98	0.91
S Lee	0.75	0.65	0.64	0.77
Y Chen	0.84	0.60	0.83	0.86
Mean	0.78	0.67	0.74	0.81
StdDev	0.10	0.13	0.14	0.08

Paper # 2 —— SVM

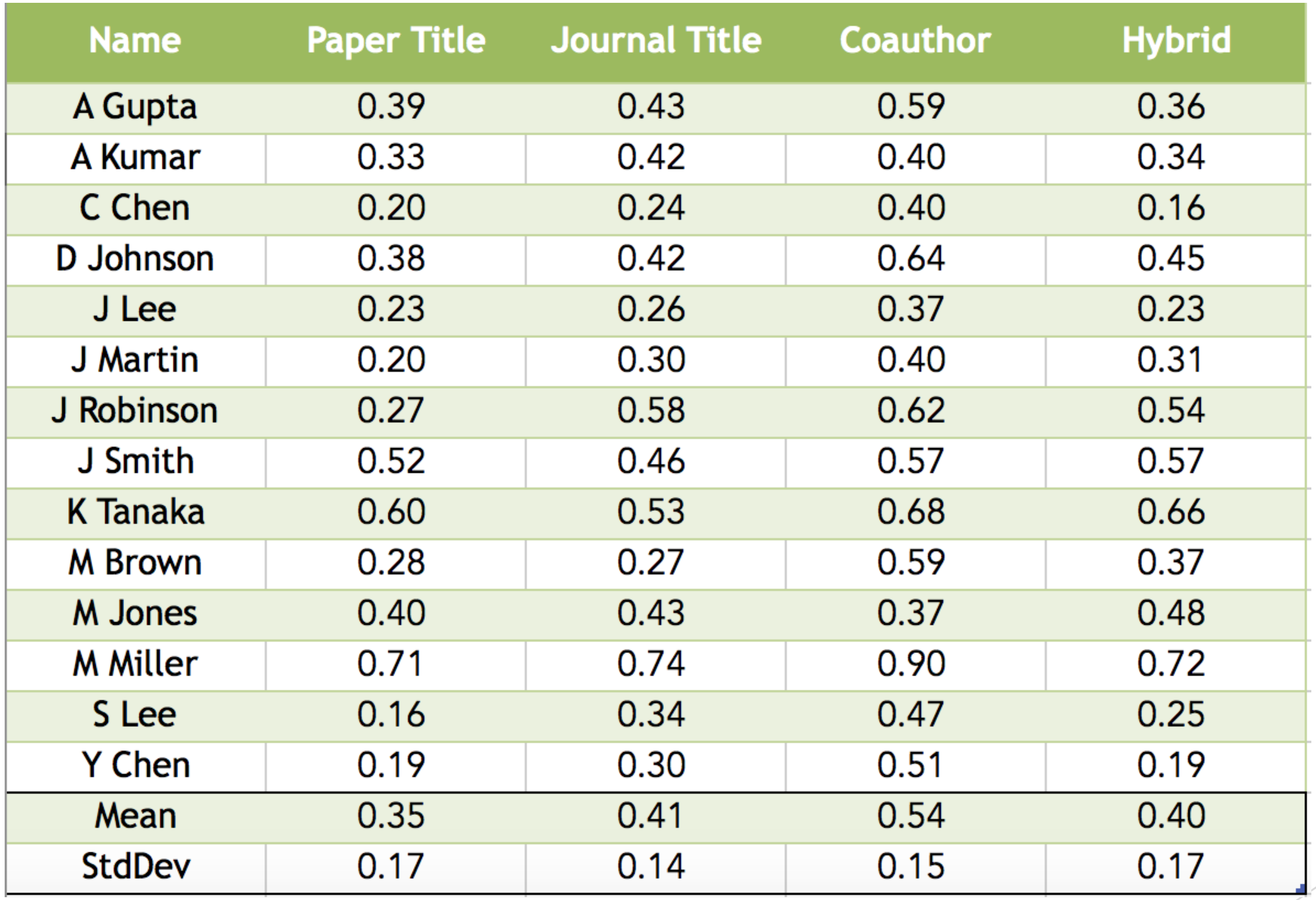
- Evaluation(recall)



Name	Paper Title	Journal Title	Coauthor	Hybrid
A Gupta	0.39	0.43	0.59	0.36
A Kumar	0.33	0.42	0.40	0.34
C Chen	0.20	0.24	0.40	0.16
D Johnson	0.38	0.42	0.64	0.45
J Lee	0.23	0.26	0.37	0.23
J Martin	0.20	0.30	0.40	0.31
J Robinson	0.27	0.58	0.62	0.54
J Smith	0.52	0.46	0.57	0.57
K Tanaka	0.60	0.53	0.68	0.66
M Brown	0.28	0.27	0.59	0.37
M Jones	0.40	0.43	0.37	0.48
M Miller	0.71	0.74	0.90	0.72
S Lee	0.16	0.34	0.47	0.25
Y Chen	0.19	0.30	0.51	0.19
Mean	0.35	0.41	0.54	0.40
StdDev	0.17	0.14	0.15	0.17

Paper # 2 —— SVM

- Evaluation(f1: 2*harmonic mean of presicion and recall)



Name	Paper Title	Journal Title	Coauthor	Hybrid
A Gupta	0.39	0.43	0.59	0.36
A Kumar	0.33	0.42	0.40	0.34
C Chen	0.20	0.24	0.40	0.16
D Johnson	0.38	0.42	0.64	0.45
J Lee	0.23	0.26	0.37	0.23
J Martin	0.20	0.30	0.40	0.31
J Robinson	0.27	0.58	0.62	0.54
J Smith	0.52	0.46	0.57	0.57
K Tanaka	0.60	0.53	0.68	0.66
M Brown	0.28	0.27	0.59	0.37
M Jones	0.40	0.43	0.37	0.48
M Miller	0.71	0.74	0.90	0.72
S Lee	0.16	0.34	0.47	0.25
Y Chen	0.19	0.30	0.51	0.19
Mean	0.35	0.41	0.54	0.40
StdDev	0.17	0.14	0.15	0.17

Paper # 6 —— HMRF(Semi- supervised framework)

- HMRF(Hidden Markov Random Fields):

$$c_l(p_i, p_j) = \begin{cases} 1 & \text{if } p_i \text{ and } p_j \text{ satisfy the constraint } c_l \\ 0 & \text{otherwise} \end{cases}$$

- Constraint selection(define w_i):

  C2(CoAuthor): Two papers have a secondary author with the same name

  C6(τ-CoAuthor): Gives deeper coauthorship within papers

define w_2=0.7, w_6=0.7^{τ}

Paper # 6 —— EM algorithm with HMRF

- Initialization:
- Define distance function:

$$D(x_i,x_j)=1-\frac{x_i^T \mathbf{A} x_j}{\| x_i \|_{\mathbf{A}} \| x_j \|_{\mathbf{A}}}\text{, where } \| x_i \|_{\mathbf{A}}=\sqrt{x_i^T \mathbf{A} x_j}$$

- E-step

$$f(y_h,x_i)=\sum_{i} D(x_i,y_h)+\sum_{i,j\neq i}\{D(x_i,x_j)\sum_{c_k\in C}[w_k c_k(x_i,x_j)]\}$$

- M-step: update the diagonal elements of matrix A

$$a_{mm}=a_{mm}+\eta\frac{\partial f_{obj}}{\partial a_{mm}}$$

Paper # 6 —— Our attempt

- Making attempts to vectorize each loop to reduce the running time (Our EM step gives a much faster speed after vectorization)
- Using permutation to test the clustering sequence which gives the most accurate result(still kind of slow)

Paper # 6 —— Evaluation

- So sorry that we’ve finished the clustering step but were not able to produce the evaluation result report before due, but we will finish that no matter how.