Finding a Team of Experts in Social Networks

T. Lappas, K. Liu, E. Terzi
Knowledge Discovery and Data Mining 2009

Presented by Alex Klibisz

Team Formation Problem

Given:

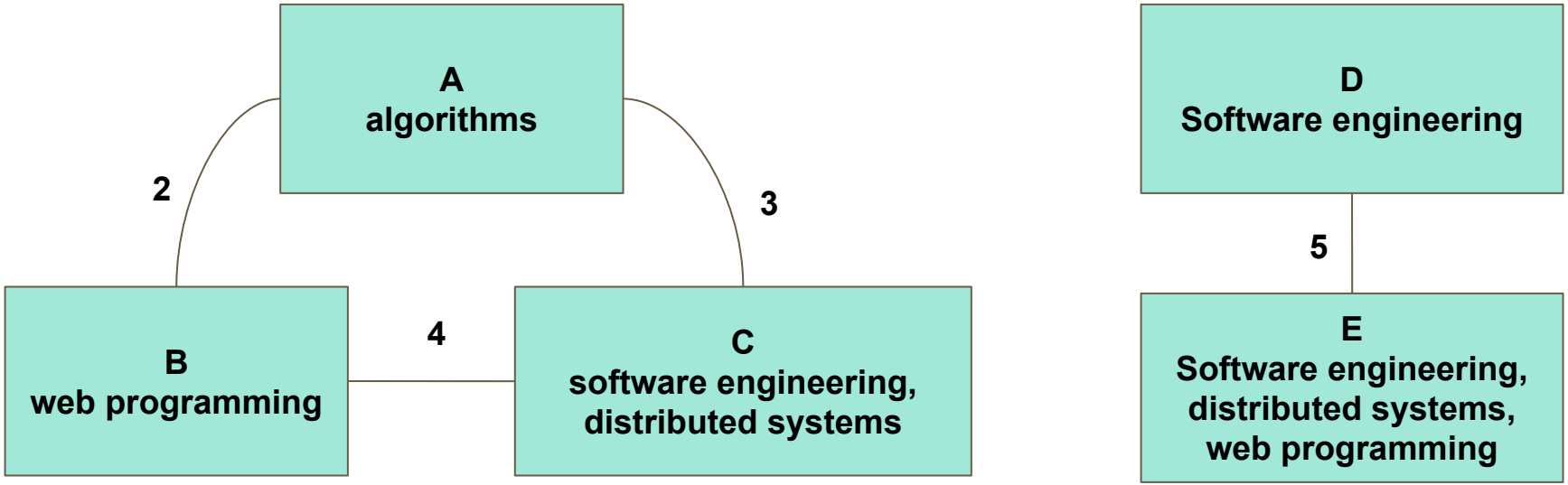
- Task requiring a set of skills
- Set of individuals
- Skills possessed by each individual
- Graph of communication cost between pairs of individuals
    - Different departments, languages, time zones, etc.

Find:

- A subset of individuals containing all required skills with minimized communication cost

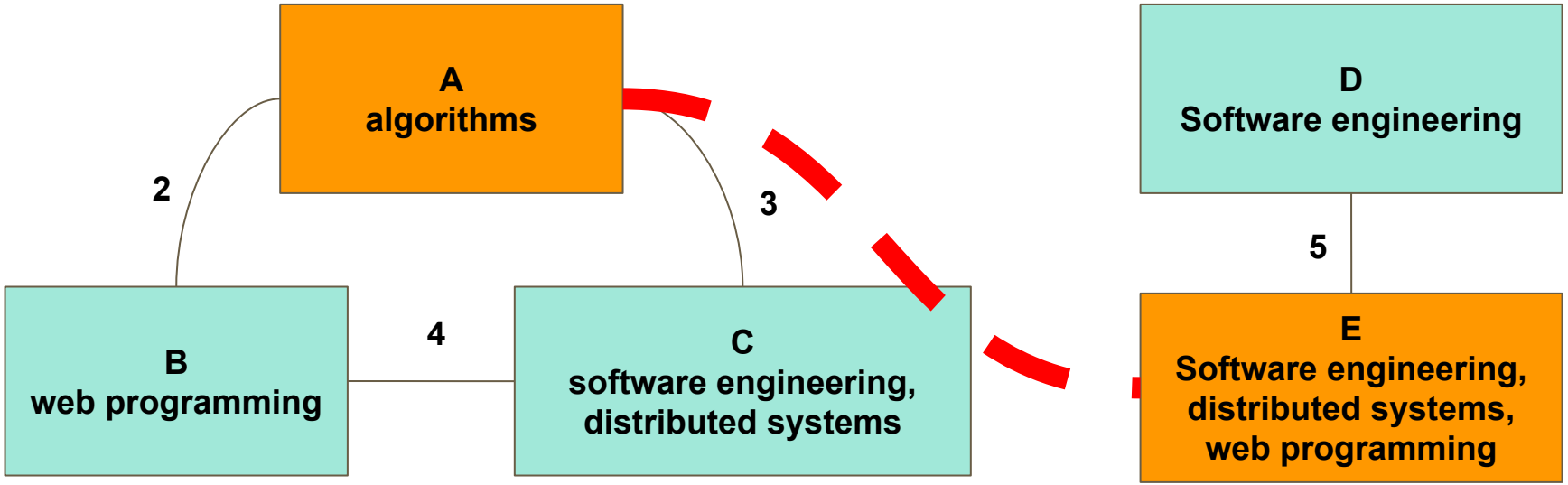
Team Formation Example

T = { algorithms, software engineering, distributed systems, web programming }



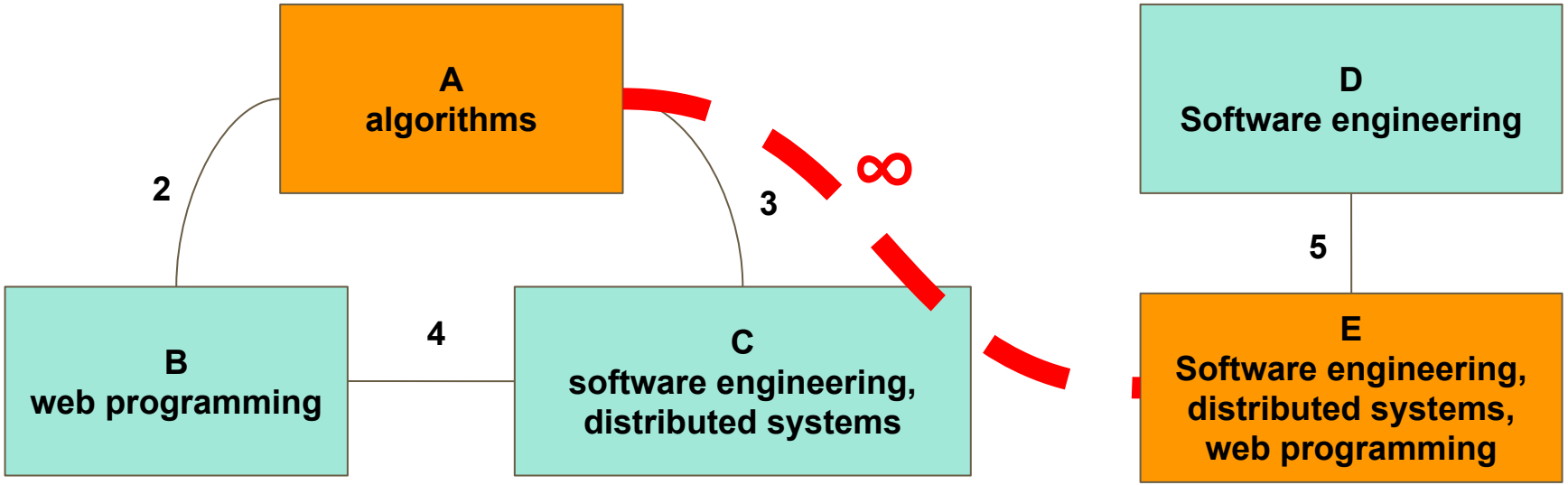
Team Formation Example

T = { algorithms, software engineering, distributed systems, web programming }



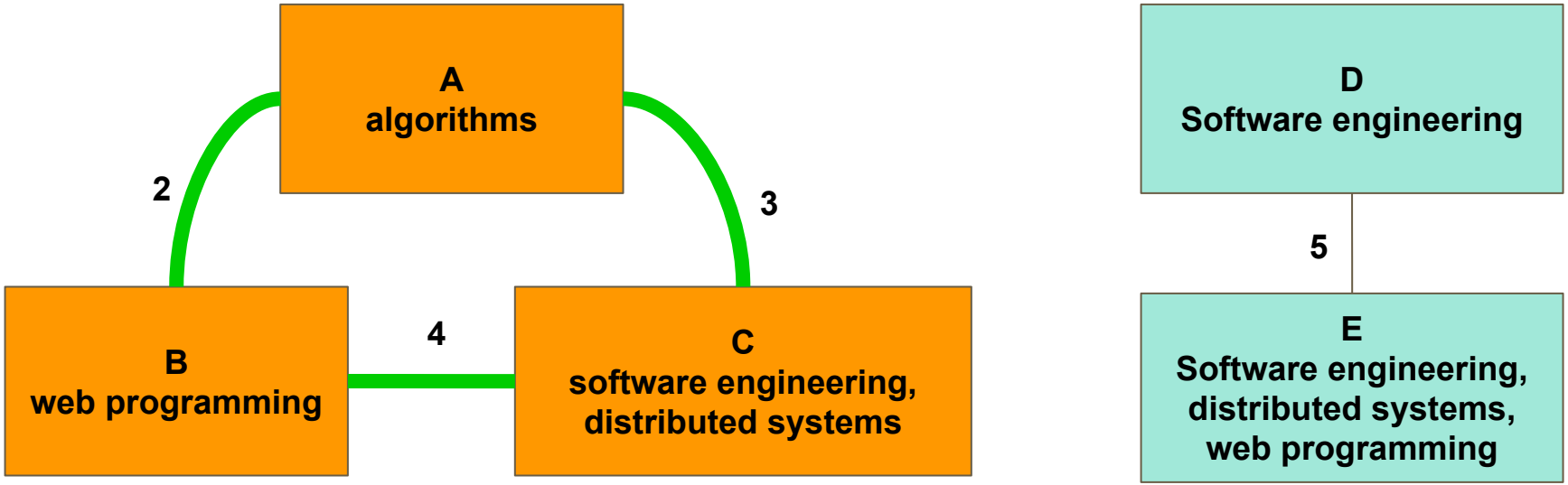
Team Formation Example

T = { algorithms, software engineering, distributed systems, web programming }



Team Formation Example

T = { algorithms, software engineering, distributed systems, web programming }



Related Work

- Match-making optimization
- Few have considered social graphs
- This work should be considered complementary

Terminology

- *Cover*: subset of all desired skills for which at least one individual possesses each skill.

- *Support Set:* set of individuals that posses a particular skill.

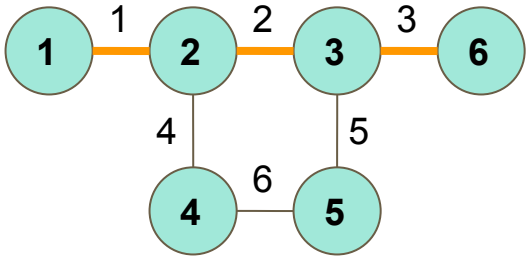
- *Distance Function:* returns weight of connection between two individuals or between one individual and the closest individual in a support set.

Problem Variations

- Different ways to measure and minimize *Communication Cost*
- Diameter-TF
- Minimum Spanning Tree TF

Diameter-TF

- Communication Cost = diameter of the graph of selected individuals
- NP-Complete: reduction of Multiple Choice Cover Problem
- NP-Hard: when distance function is a metric



Diameter = 6

MST-TF

- Communication Cost = cost of Minimum Spanning Tree over selected individuals
- NP-Complete, NP-Hard: reduction of Group-Steiner Tree Problem

Diameter-TF: Solution, *RarestFirst*

1. **Compute the support for every skill, a, required by task T:**
2. **Pick the rarest skill, a_{rare}**
3. **For every person in support(a_{rare}), connect the person to the closest support of every other skill**

Details

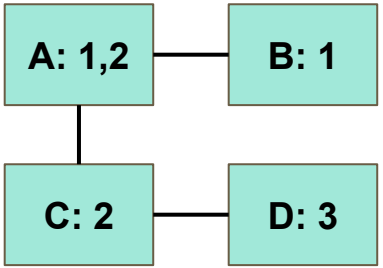
- Assume pre-computed shortest paths between pairs
- Time complexity = O(|S(a_{rare})| x n) for n individuals
- **Worst-case = $O(n^2)$**

Diameter-TF Example

Given:

X = { A, B, C, D }

T = { 1, 2, 3}



Apply RarestFirst:

a_{rare} = 3, s(3) = { D }

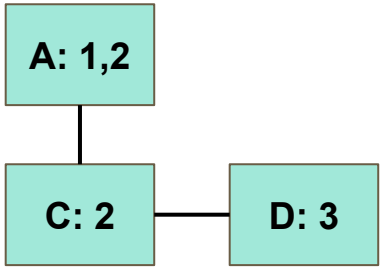
D → s(1): { A, D }

D → s(2): { A, C, D }

Result

X' = { A, C, D }

CC-D(X') = 1



MST-TF: Solution 1, *CoverSteiner*

For set of individuals X, task T:

1. X_0 = GreedyCover(X, T)
2. X′ = SteinerTree(G, X_0)

GreedyCover: Find a set of individuals that covers all skills in T

1. While covered skills != T
    a. Add individual with most uncovered skills

SteinerTree: Find minimum cost spanning tree for X_0

1. While X′ != X_0
    a. Find single node v* from X_0 that has min distance to X′

CoverSteiner cont.

Time Complexity

- GreedyCover = $O(|T| \times |X|) = O(mn)$;
  - m is the number of skills in T, n is the number of total individuals
- SteinerTree = $O(|X_0| \times |E|)$
  - E are the edges connecting individuals
- **Worst-case = $O(n^3)$**

Flaw

- Step 1 completely ignores the graph structure, leading to high communication costs

MST-TF: Solution 2, *EnhancedSteiner*

1. H, Y = EnhanceGraph(G, T)
2. X_H = SteinerTree$(H, \{Y_1, \ldots, Y_k\})$
3. $X' = X_H \setminus \{Y_1, \ldots, Y_k\}$

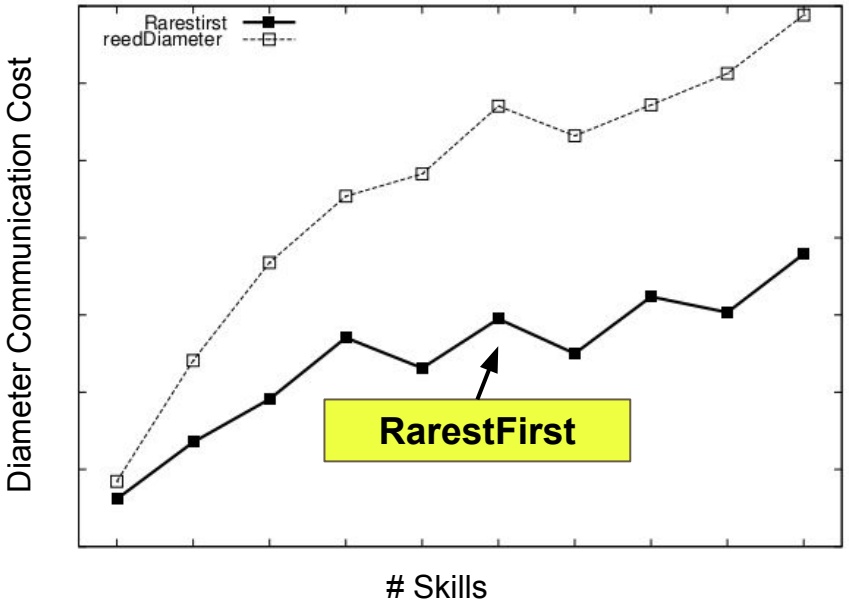
EnhanceGraph: Add artificial nodes, edges to minimize SteinerTree communication cost

1. For every skill a_j in T
   a. Create an additional node Y_j
   b. Connect node Y_j to all individuals with skill a_j with large weight
   c. All nodes with skill a_j are represented as a *clique*

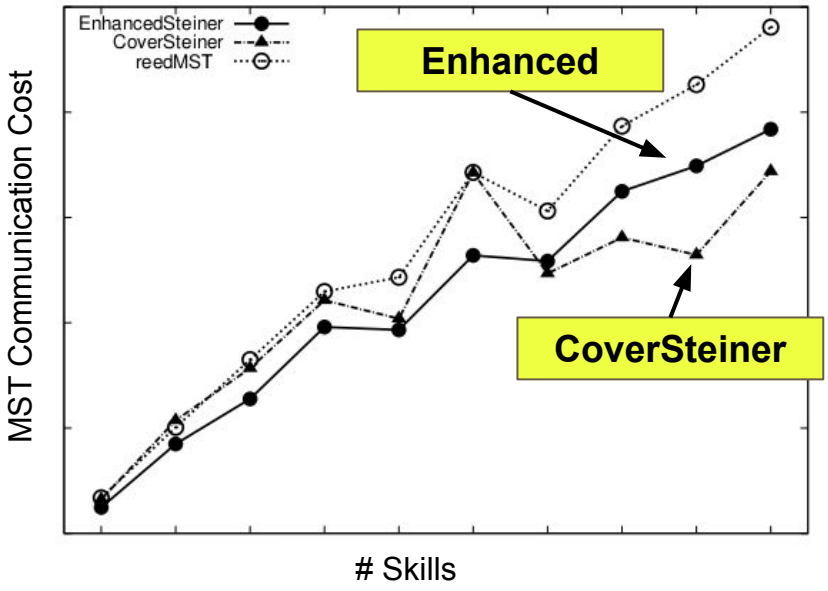
Experimental Evaluation

- DBLP: papers in *databases, data mining, artificial intelligence, theory*
- Skills derived from common terms in paper titles
- Communication weights determined by co-authorship
- 5509 individuals, 1792 skills
- Tasks generated with 2 to 20 skills
- Average over 100 combinations for final values

Performance: Communication Cost

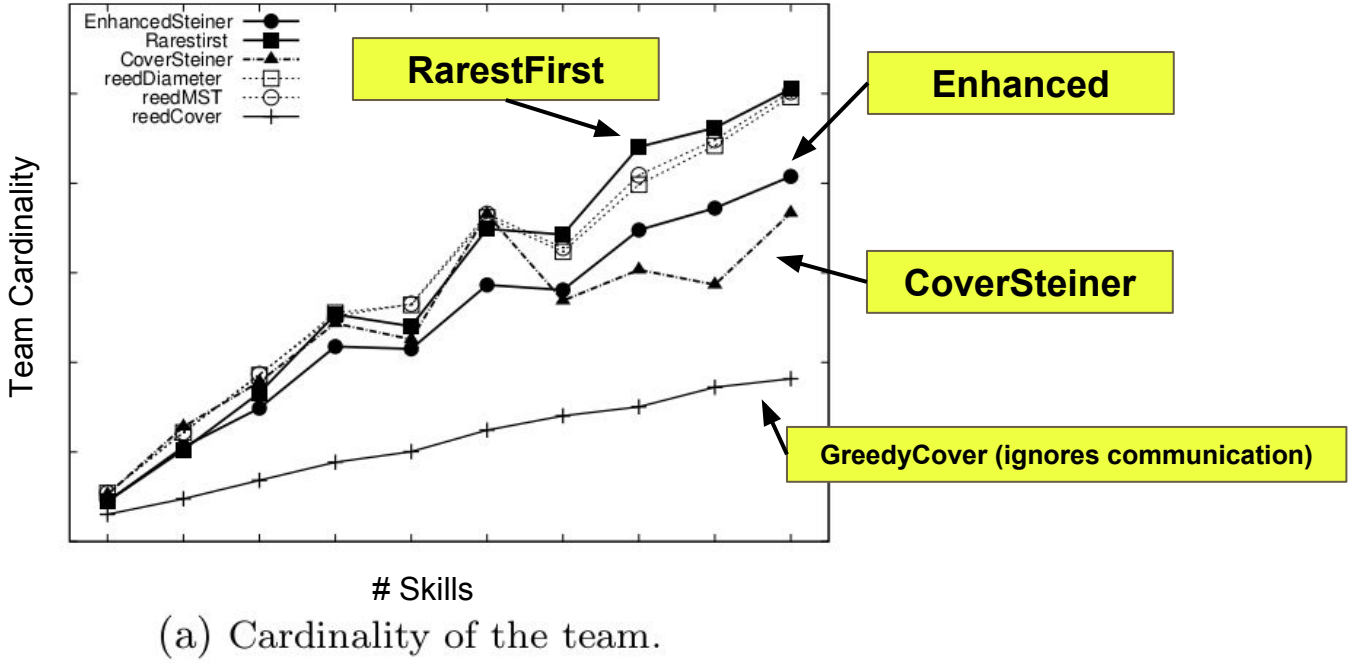


(a) Cc-R cost



(b) Cc-Mst cost

Performance: Team Cardinality



(a) Cardinality of the team.

Team Formation Results

Rank	Actual authors	`RarestFirst` result	`EnhancedSteiner` result
1	S. Brin, L. Page	**Paolo Ferragina, Patrick Valduriez, H. V. Jagadish, Alon Y. Levy, Daniela Florescu** Divesh Srivastava, S. Muthukrishnan	**P. Ferragina ,J. Han, H. V. Jagadish, Kevin Chen-Chuan Chang, A. Gulli**, S. Muthukrishnan, Laks V. S. Lakshmanan
2	R. Agrawal, R. Srikant	**R. Agrawal**	**Philip S. Yu**
3	R. Agrawal, T. Imielinski, A. N. Swami	**Philip S. Yu**	**Wei Wang, Philip S. Yu**
4	T. Joachims	**Wei-Ying Ma, Gui-Rong Xue, H. Liu, J. Han, H. Lu, Z. Chen**, Q.Yang, H. Cheng	**J. Han, H. Lu, Wei-Ying Ma, Z. Chen, H. Liu, Gui-Rong Xue**, Q. Yang
5	J. Lafferty, F. Pereira, A. McCallum	**A. McCallum**	**A. McCallum**
6	J. Han, J. Pei, Y. Yin	**F. Bonchi**	**A. Gionis, H. Mannila, R. Motwani**
7	E. Rahm, P. A. Bernstein	**C. Bettini, R. Agrawal, Kevin Chen-Chuan Chang**, T. Imielinski, H. Garcia-Molina, D. Barbara, S. Jajodia	**C. Bettini, P. A. Bernstein**, H. Garcia-Molina, S. Jajodia, D. Maier, D. Barbara
8	R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan	**D. Gunopulos, R. Agrawal**	**R. Agrawal, D. Gunopulos**
9	B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom	**M. T. Ozsu**	**H. V. Jagadish, D. Srivastava**
10	J. Chen, D. J. DeWitt, F. Tian, Y. Wang	**Donald Kossmann, David J. DeWitt, Michael J. Franklin, Michael J. Carey**	**M. J. Carey, M. J. Franklin, D. Kossmann, D. J. DeWitt**

Thoughts

- Limited Dataset
  - Does using titles for skills actually reflect the skills?
  - Not all authors on a paper have all of the skills
- Problem variation
  - Quantify the quality of a person's skill
- Replicate with Kaggle team data:
  - Create a "skills" feature for each competition and individual (ex: classification, regression, computer vision)
  - Compare generated teams to actual teams