PROBLEM
To develop a star schema, existing approaches analyze the attributes of interesting business entities.
• Entities with numerical measure attributes are assumed to be the candidates of facts
• Entities with non-numerical and descriptive attributes are assumed to be the candidates of dimensions.

Hence, these approaches are qualitative in nature, and focus only on the semantics of ERD.

OUR SOLUTION
We focus on the structure of the ERD. The novel features of SAMSTAR are:
(1) the use of the notion of Connection Topology Value (CTV) in identifying the candidates of facts and dimensions and
(2) the use of Annotated Dimension Design Patterns (A-DDP) as well as WordNet to extend the list of dimensions.

1. Connection Topology Value (CTV)

\[ CTV(e) = 1^n + 0.8 \sum \text{CTV(Node(i))] \]

where \( i \) represents an entity having a direct M:1 relationship with e.

For Fig. 1, CTV is calculated in the following manner:

\[
\text{weight}_d = 100\%; \text{weight}_i = 80\% \\
\text{The CTV for each entity is:} \\
\text{CTV(H)} = 1^1 + 0.8 * 0 = 0 \\
\text{CTV(F)} = 0 \\
\text{CTV(G)} = 0 \\
\text{CTV(E)} = 1^1 + 0.8 * \text{CTV(H)} = 1 \\
\text{CTV(B)} = 1^1 + 0.8 * \text{CTV(E)} = 1.8 \\
\text{CTV(C)} = 1^2 + 0.8 * \text{CTV(G) + CTV(F)} = 2 \\
\text{CTV(D)} = 1^1 + 0.8 * \text{CTV(C)} = 1 + 0.8 * 2 = 2.6 \\
\text{CTV(A)} = 1^2 + 0.8 * (\text{CTV(B) + CTV(C)}) = 2 + 0.8 * (1.8+2) = 5.04
\]

2. Annotated Dimension Design Patterns (A-DDP)
We have referred to four sources and have instantiated the six classes of DDP to produce a list of 131 commonly used dimension entities. We refer to these entities as Annotated DDP (A-DDPs). These entities are frequently used in the business processes. Examples include account, activity, agent, aircraft, airport, etc.

ALGORITHM

1. Pre-process the input ERD.
2. Store Entities and Relationships.
3. Let user choose weighting factors for direct and indirect relationships.
4. Calculate the CTV for all entities.
5. Calculate the threshold value, Th, for CTV.
6. Identify entities having CTV higher than the threshold Th.
   These are the candidates for fact tables.
7. Decide and shortlist the fact entities.
8. For each fact entity, perform the following steps:
   (i) Identify the entities having direct M:1 link with a fact entity.
   (ii) Identify entities having indirect M:1 link with the fact entity.
   - Out of these entities, identify synonyms of entity names from WordNet.
   - Extract the terms which match the DDP entity list or the A-DDP List.
   (iii) Combine the results to Steps 8(i) and 8(ii) to prepare a list of candidate dimensions for a given fact.
   (iv) Add time dimension to this list.
9. Decide the dimension entities.
10. Let the user post-process the Star Schemas
11. Generate the star schema(s).

RESULTS
Schemas generated by SAMSTAR are similar to those generated by the manual steps in case study papers.
SAMSTAR generated star schemas are the superset of the ones generated manually in the paper using the same ER diagrams, user needs and business goals.
This shows our schemas are inclusive of all possible facts and dimensions. This gives the designer a helpful aid to prune the schema as per the business and user requirements.

CONTRIBUTION
• A universal method to generate star schema(s) in that we have used generalized DDPs and WordNet to identify dimensions of a fact table.
• Quantitative in nature in that we analyze the structure of an input ERD.
• Identifies a set of fact candidates from a large and complex ERD.
• Automatable up to a large extent; simplifies the work of experienced designers; and gives a smooth head-start to novices.

www.ischool.drexel.edu