

# Generator for Noisy Reference Data with Co-occurrence Relationships

Indrajit Bhattacharya (indrajit@cs.umd.edu)

Lise Getoor (getoor@cs.umd.edu)

January 30, 2007

We found synthetically generated noisy reference data to be very helpful in our evaluations. Experiments on synthetic data enabled us to reason beyond specific datasets, understand the impact of different structural properties of the data on collective resolution, and also to empirically verify our performance analysis for relational clustering in general.

We designed a synthetic data generator with a two-stage generation process. In the first stage, we create the collaboration graph among the underlying entities and the entity attributes. In the second, we generate observed co-occurrence relations from this collaboration graph. A high level description of the generative process is shown in Figure 1. Next, we describe the two stages of the generation process in greater detail.

The graph creation stage, in turn, has two sub-stages. First, we create the domain entities and their attributes and then add relationships between them. For creating entities, we control the number of entities and the ambiguity of their attributes. We create  $N$  entities and their attributes one after another. For simplicity and without losing generality, each entity  $e$  has a single floating point attribute  $e.x$ , instead of a character string. A parameter  $p_a$  controls the ambiguity of the entity attributes; with probability  $p_a$  the attribute of a new entity is chosen from values that are already in use by existing entities. Then  $M$  binary relationships are added between the created entities. As with the attributes, there is a parameter controlling the ambiguity of the relationships. For each binary relationship  $(e_i, e_j)$ , first  $e_i$  is chosen randomly and then  $e_j$  is sampled so that  $(e_i, e_j)$  is an ambiguous relationship with probability  $p_a^R$ .

Before describing the process of generating co-occurrence relationships from the graph, let us consider in a little more detail the issue of attribute ambiguity. What finally needs to be controlled is the ambiguity of the reference attributes. While these depend on the entity attributes, they are not completely determined by entities. Taking the example of names, two people who have names ‘John Michael Smyth’ and ‘James Daniel Smith’ can still be ambiguous in terms of their observed names in the data depending on the generation process of observed names. In other words, attribute ambiguity of the references depends both on the separation between entity attributes and the dispersion created

by the generation process. We make the assumption that for an entity  $e$  with attribute  $e.x$ , its references are generated from a Gaussian distribution with mean  $x$  and variance 1.0. So, with very high probability, any reference attribute generated from  $e.x$  will be in the range  $[e.x - 3, e.x + 3]$ . So this range in the attribute domain is considered to be ‘occupied’ by entity  $e$ . Any entity has an ambiguous attribute if its occupied range intersects with that of another entity.

Now we come to the generation of co-occurrence relationships from the entity collaboration graph. In this stage,  $R$  co-occurrence relationships or hyper-edges are generated, each with its own references. For each hyper-edge  $\langle r_i, r_{i1}, \dots, r_{ik} \rangle$ , two aspects need to be controlled — how many references and which references should be included in this hyper-edge. This is done as follows. First, we sample an entity  $e_i$  which serves the initiator entity for this hyper-edge. Then other entities  $e_{ij}$  for this hyper-edge are repeatedly sampled (without replacement) from the neighbors of the initiator entity  $e_i$ . The size of the hyper-edge is determined using a parameter  $p_c$ . The sampling step for a hyper-edge is terminated with probability  $p_c$  after each selection  $e_{ij}$ . The process is also terminated when the neighbors of the initiator entity are exhausted. Finally, references  $r_{ij}$  need to be generated from each of the selected entities  $e_{ij}$ . This is done for each entity  $e$  by sampling from its Gaussian distribution  $\mathcal{N}(e.x, 1)$ .

- Creation Stage
1. Repeat N times
    2. Create random attribute  $x$  with ambiguity  $p_a$
    3. Create entity  $e$  with attribute  $x$
  4. Repeat M times
    5. Choose entity  $e_i$  randomly
    6. Choose entity  $e_j$  with prob  $p_a^R$  of an ambiguous relationship  $(e_i, e_j)$
    7. Set  $e_i = Nbr(e_j)$  and  $e_j = Nbr(e_i)$
- Generation Stage
8. Repeat R times
    9. Randomly choose entity  $e$
    10. Generate reference  $r$  using  $\mathcal{N}(e.x, 1)$
    11. Initialize hyper-edge  $h = \langle r \rangle$
    12. Repeat with probability  $p_c$ 
      13. Randomly choose  $e_j$  from  $Nbr(e)$  without replacement
      14. Generate reference  $r_j$  using  $\mathcal{N}(e_j.x, 1)$
      15. Add  $r_j$  hyper-edge  $h$
    16. Output hyper-edge  $h$

Figure 1: High-level description of synthetic data generation algorithm