

# Automatically Learning Author Relationships from Bibliography Data

Akash Kushal

*kushal@cs.uiuc.edu*

Department of Computer Science  
University of Illinois, Urbana-Champaign

## Abstract

This work presents an approach to automatically learn the academic relationships between authors from a bibliography database. A probabilistic generative model framework is proposed to model the relationships. The references in the bibliography are the observed variables in the model and the relationships between the authors form the set of hidden variables. A variational EM algorithm is used to simultaneously learn the parameters of the model as well as the hidden author relationships. The learning process does not require the use of any labeled training data.

The proposed approach has been implemented on the DBLP Computer Science Bibliography Database [6]. The results have been made available online at <http://www-cvr.ai.uiuc.edu/~kushal/courses/cs512/advisor.php>. A human labeled dataset of over a hundred PhD students of ten different advisors in the computer science department was used to test the performance of the method. The proposed generative model supports only advisor-advisee relations between authors however extensions to modeling collaborators are discussed in the future work section.

## 1 Introduction

Considering one's doctoral advisor as one's academic parent, we can define an academic genealogy of researchers which encodes the academic ancestors and descendants of a particular set of researchers. We may construct a graph with the authors as the vertices and add a directed edge from every advisee to his advisor. If we allow a student to have a single advisor, this structure becomes a directed tree since the advisor of every student must graduate before the student. This tree is called the academic genealogy tree and is very similar to the classical genealogy tree defined by considering parent-child relations.

Many projects have been set up to maintain such information for various research fields. These include the Mathematics Genealogy Project [3], the Computer Engineering Academic Genealogy [9], the AI Genealogy Project [5] and the Software Engineering Academic Genealogy [8]. However, all of these projects

rely on manually collecting the academic genealogy data which makes them quite costly. The proposed approach seeks to automatically extract this information by mining bibliographic databases.

There is a vast amount of bibliographic information available online. Perhaps the largest of these sources in the computer science domain is the DBLP Computer Science Bibliography Database [6] maintained by Michael Ley. The DBLP bibliography data consists of a huge XML file that contains information of over 800,000 distinct publications of about 500,000 distinct<sup>1</sup> authors. Each publication is associated with an ordered list of authors, the year of publication and other information such as the name of the conference or journal where the publication appeared.

One would expect that before a student graduates from his advisors group, the student would collaborate with the advisor on a large fraction of his or her publications. The student would also be likely to include other students from his or her advisor’s group during that period as co-authors in the publications. In this work, we present an approach to systematically combine these cues to automatically learn advisor-advisee relationships from bibliography data. Apart from the advisor-advisee relationships, our model also predicts the approximate graduation year of each of the different authors in the database. Since, acquiring labeled training data is costly, we use a generative model based approach that does not require any labeled training data during the learning process. The advisor-advisee relations and the graduation years (and two other specific points in the research time-line) for the various authors are considered as hidden variables in the model. The references in the bibliography database form the set of observed variables. Variational EM [4] is used to simultaneously learn the parameters of the generative model as well as the posterior distribution over the advisor relations and graduation dates for all the authors.

The required notation and a mathematical formulation of the learning problem is provided in section 2. Section 3 describes the details of the proposed generative model. The details of the variational EM procedure for learning the model parameters and the maximum-a-posteriori (MAP) inference procedure are provided in section 4. The results obtained by implementing the proposed approach on the DBLP database are presented in section 5. Finally, section 6 discusses further extensions of the method and provides a conclusion for the work.

## 2 Problem Formulation

As briefly mentioned in the previous section, every author  $\mathcal{A}$  in the database is associated with a set of hidden variables:

- $\mathcal{A}.adv$  :  $\mathcal{A}$ ’s advisor during his student years.

---

<sup>1</sup>Since a single author may publish under different names and many distinct authors may have a single name the actual number of authors may be different. An entity deduplication method like [1] and an object distinction method like [10] could be used to preprocess the data and reduce these problems.

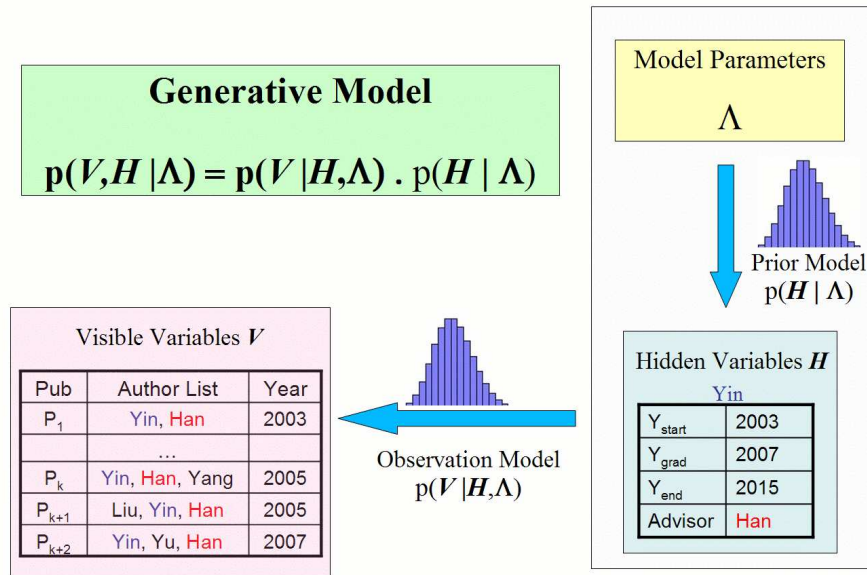


Figure 1: The Generative Model consists of two components: (1) The prior model  $p(\mathcal{H}|\Lambda)$  and (2) the observation model  $p(\mathcal{V}|\mathcal{H},\Lambda)$ . The hidden variables consist of the advisor and the years  $\mathcal{A}.y_{start}$ ,  $\mathcal{A}.y_{grad}$  and  $\mathcal{A}.y_{end}$ . Note that the year  $\mathcal{A}.y_{start}$  may be before the year corresponding to the first publication of  $\mathcal{A}$  and the year  $\mathcal{A}.y_{end}$  may be after the last year of publications. In fact, as shown in this example, the model may even predict a value for  $\mathcal{A}.y_{end}$  that lies in the future. The visible variables consist of the publications in DBLP.

- $\mathcal{A}.y_{start}$  : the year when  $\mathcal{A}$  starts research.
- $\mathcal{A}.y_{grad}$  : the year when  $\mathcal{A}$  graduates from his advisors group.
- $\mathcal{A}.y_{end}$  : the year when  $\mathcal{A}$  stops active research.

During the time from  $\mathcal{A}.y_{start}$  to  $\mathcal{A}.y_{end}$ , the author may publish various papers. The list of publication records associated with author  $\mathcal{A}$  is  $\mathcal{A}.P = \{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_{n_{\mathcal{A}}}\}$  which form the visible variables. Each publication record  $\mathcal{P} = (\{\mathcal{A}_{\mathcal{P}}(1), \mathcal{A}_{\mathcal{P}}(2) \dots \mathcal{A}_{\mathcal{P}}(k_{\mathcal{P}})\}, y_{\mathcal{P}})$  consists of a sequence of co-authors and a publication year. We denote by  $\mathcal{H}$  (resp.  $\mathcal{V}$ ) the set of all the hidden (resp. visible) variables in the model. The joint probability distribution  $p(\mathcal{H}, \mathcal{V}|\Lambda)$  over all the hidden variables  $\mathcal{H}$  and the visible variables  $\mathcal{V}$  is modeled in a **generative model** framework. Here  $\Lambda$  denotes the set of all the parameters in the generative model. Figure 1 shows the different components of the generative model diagrammatically.

In a typical generative model framework like ours, we are faced with two tasks:

- **Model Learning:** Learning consists of computing the parameters  $\Lambda^*$  that maximize the likelihood of observing the visible variables  $\mathcal{V}$  after marginalizing over the hidden variables  $\mathcal{H}$ . That is, we would like to compute:

$$\Lambda^* = \arg \max_{\Lambda} p(\mathcal{V}|\Lambda) = \arg \max_{\Lambda} \left[ \sum_{\mathcal{H}} p(\mathcal{H}, \mathcal{V}|\Lambda) \right] \quad (1)$$

- **Inference of Hidden Variables:** The inference process seeks to compute the posterior probability distribution  $p(\mathcal{H}|\mathcal{V}, \Lambda)$  over the hidden variables  $\mathcal{H}$  in the model given the observed data  $\mathcal{V}$  and the learned parameters  $\Lambda$ . In many cases, we are only interested in the mode of the posterior distribution  $\mathcal{H}^*$  which corresponds to the most probable values for the hidden variables given the observed data and the parameters. The Maximum A Posteriori (or MAP) estimate  $\mathcal{H}^*$  can be written as:

$$\mathcal{H}^* = \arg \max_{\mathcal{H}} p(\mathcal{H}|\mathcal{V}, \Lambda) \quad (2)$$

## 2.1 Generative versus Discriminative Models

The generative approach employed in our work is different from a typical discriminative approach which would directly model the posterior probability  $p(\mathcal{H}|\mathcal{V}, \Lambda)$  of the hidden variables  $\mathcal{H}$  given the visible variables  $\mathcal{V}$  and model parameters  $\Lambda$ . The reason for choosing a generative model in our case is that we want to conduct the learning without the use of labeled training data. This is not possible if we model  $p(\mathcal{H}|\mathcal{V}, \Lambda)$  directly. Without the use of labeled training data, the only observed distribution is  $p(\mathcal{V})$  and it does not provide any information about the distribution to be learned  $p(\mathcal{H}|\mathcal{V}, \Lambda)$ . Hence, discriminative models require the use of labeled training data during learning.

## 3 Generative Model Description

The joint probability distribution  $p(\mathcal{H}, \mathcal{V}|\Lambda)$  may be factorized as a product of the prior model  $p(\mathcal{H}|\Lambda)$  and the observation model  $p(\mathcal{V}|\mathcal{H}, \Lambda)$ .

$$p(\mathcal{H}, \mathcal{V}|\Lambda) = p(\mathcal{V}|\mathcal{H}, \Lambda)p(\mathcal{H}|\Lambda)$$

### 3.1 Prior Distribution Model $p(\mathcal{H}|\Lambda)$

We next describe the components of the prior distribution  $p(\mathcal{H}|\Lambda)$  defined over the hidden variables  $\mathcal{H}$ . The variables in  $\mathcal{H}$  may be broken down into two groups: 1)  $\mathcal{V}$ , the set of variables  $\mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \mathcal{A}.y_{end}$  for all the authors, and 2)  $\mathcal{M}$ , the set of the variables  $\mathcal{A}.adv$  for all the authors. Hence, we may decompose  $p(\mathcal{H}|\Lambda)$  as

$$p(\mathcal{H} = \mathcal{Y} \cup \mathcal{M} | \Lambda) = p(\mathcal{Y} | \Lambda) p(\mathcal{M} | \mathcal{Y}, \Lambda)$$

We assume the variables in  $\mathcal{Y}$  corresponding to the different authors are generated independently. That is,

$$p(\mathcal{Y} | \Lambda) = \prod_{\mathcal{A} \in DBLP} p(\mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \mathcal{A}.y_{end} | \Lambda)$$

The distribution  $p(\mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \mathcal{A}.y_{end})$  may be decomposed as:

$$\begin{aligned} p(\mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \mathcal{A}.y_{end} | \Lambda) &= p(\mathcal{A}.y_{start} | \Lambda) \times p(\mathcal{A}.y_{grad} | \mathcal{A}.y_{start}, \Lambda) \\ &\quad \times p(\mathcal{A}.y_{end} | \mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \Lambda) \end{aligned}$$

where,

1. The year  $\mathcal{A}.y_{start}$  in which  $\mathcal{A}$  starts his research is assumed to come from a uniform distribution with fixed range  $M$  (say from 1907 to 2007). It turns out that the final inference result is completely independent of the value of  $M$ . It is included just for the purpose of properly defining the joint distribution  $P(\mathcal{H}, \mathcal{Y} | \Lambda)$ .

$$p(\mathcal{A}.y_{start}) = \frac{1}{M}$$

2. The distribution of the number of student years ( $\mathcal{A}.y_{grad} - \mathcal{A}.y_{start}$ ) of  $\mathcal{A}$  is assumed to be a Poisson distribution with mean  $\lambda_N$ . Hence,

$$p(\mathcal{A}.y_{grad} | \mathcal{A}.y_{start}, \Lambda) = p_{poiss}(\mathcal{A}.y_{grad} - \mathcal{A}.y_{start} | \lambda_N)$$

3. The number of active years of an author  $\mathcal{A}$  after graduation ( $\mathcal{A}.y_{end} - \mathcal{A}.y_{grad}$ ) is assumed to be a Poisson distribution with mean  $\lambda_{act}$ . Hence,

$$p(\mathcal{A}.y_{end} | \mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \Lambda) = p_{poiss}(\mathcal{A}.y_{end} - \mathcal{A}.y_{grad} | \lambda_{act})$$

We assume that the advisors of each of the authors are chosen independently given the dates in  $\mathcal{Y}$  and the model parameters  $\Lambda$ . Hence, we can write

$$p(\mathcal{M} | \mathcal{Y}, \Lambda) = \prod_{\mathcal{A} \in DBLP} p(\mathcal{A}.adv | \mathcal{Y}, \Lambda).$$

The advisor of an author  $\mathcal{A}$  may or may not be an author in the database. One of the cases when this happens is when the  $\mathcal{A}$  is a senior researcher and the papers published by  $\mathcal{A}$  during his or her student years are not part of DBLP. In such a case the hidden variable  $\mathcal{A}.adv$  is set equal to *NOT\_IN\_DB*. We model this fact by assigning a constant probability  $p_{db}$  for any author's advisor to be a part of DBLP. If the advisor is an author in DBLP, it may be (with an equal probability) any author who graduated before the student starts his research and remained active until the student graduates. Hence,

$$p(\mathcal{A}.adv|\mathcal{Y}, \Lambda) = \begin{cases} p_{ab}/|Adv(\mathcal{A}, \mathcal{Y})| & \text{if } \mathcal{A}.adv \in DBLP \\ 1 - p_{ab} & \text{if } \mathcal{A}.adv = NOT\_IN\_DB \end{cases}$$

where the set  $Adv(\mathcal{A}, \mathcal{Y})$  denotes the set of authors that may be advisors of  $\mathcal{A}$  given the information in  $\mathcal{Y}$ . That is,  $\mathcal{A}_p \in Adv(\mathcal{A}, \mathcal{Y})$  iff  $\mathcal{A}_p.Y_{grad} < \mathcal{A}.Y_{start}$  and  $\mathcal{A}_p.Y_{end} \geq \mathcal{A}.Y_{grad}$ .

The advisor-advisee links generated by the defined prior distribution form a forest with the authors as the vertices. This is because each author has at most one advisor, and the definition of  $Adv(\mathcal{A}, \mathcal{Y})$  does not allow for cycles to exist.

### 3.2 Observation Model $p(\mathcal{V}|\mathcal{H}, \Lambda)$

The observation model describes the process of generation of the observations  $\mathcal{V}$  based on the hidden variables  $\mathcal{H}$  and model parameters  $\Lambda$ . Recall that the observed variables in this case are the publications in DBLP. Each publication record  $\mathcal{P} = (\{\mathcal{A}_{\mathcal{P}}(1), \mathcal{A}_{\mathcal{P}}(2) \dots \mathcal{A}_{\mathcal{P}}(k_{\mathcal{P}})\}, y_{\mathcal{P}})$  consists of a sequence of co-authors and a publication year. The publication  $\mathcal{P}$  is assumed to be generated by the first author  $\mathcal{A}_{\mathcal{P}}(1)$  who then picks his co-authors based on his academic relations during the publication year  $y_{\mathcal{P}}$ .

We assume that each author generates (first author) papers independently of every other author. Let  $\mathcal{V}_{\mathcal{A}}$  denote the set of papers generated by  $\mathcal{A}$ . Hence,  $p(\mathcal{V}|\mathcal{H}, \Lambda)$  may be factorized as

$$p(\mathcal{V}|\mathcal{H}, \Lambda) = \prod_{\mathcal{A} \in DBLP} p(\mathcal{V}_{\mathcal{A}}|\mathcal{H}, \Lambda).$$

Let  $\mathcal{V}_{\mathcal{A}}(y)$  be the set of (first author) publications generated by  $\mathcal{A}$  in year  $y$ . We also assume that given  $\mathcal{H}$  and  $\Lambda$ , the set of publications  $\mathcal{V}_{\mathcal{A}}(y)$  generated by  $\mathcal{A}$  in an year  $y$  are independently generated for the different years. That is,

$$p(\mathcal{V}_{\mathcal{A}}|\mathcal{H}, \Lambda) = \prod_{\mathcal{A}.y_{start} \leq y \leq \mathcal{A}.y_{end}} p(\mathcal{V}_{\mathcal{A}}(y)|\mathcal{H}, \Lambda)$$

The number of publications  $n_{\mathcal{A}}(y) = |\mathcal{V}_{\mathcal{A}}(y)|$  generated in year  $y$  is distributed according to a Poisson distribution with mean  $\lambda_S$  or  $\lambda_A$  based on whether the author  $\mathcal{A}$  is a student in year  $y$  or has already graduated<sup>2</sup>. Also, given the number of publications  $n_{\mathcal{A}}(y)$  during year  $y$ , the co-authors for each publication are chosen independently. Hence, we may write

$$p(\mathcal{V}_{\mathcal{A}}(y)|\mathcal{H}, \Lambda) = p(n_{\mathcal{A}}(y)|\mathcal{H}, \Lambda) \prod_{\mathcal{P} \in \mathcal{V}_{\mathcal{A}}(y)} p(\mathcal{P}|\mathcal{H}, \Lambda)$$

$$p(n_{\mathcal{A}}(y)|\mathcal{H}, \Lambda) = \begin{cases} p_{poiss}(n_{\mathcal{A}}(y)|\lambda_S) & \text{if } y \leq \mathcal{A}.y_{grad} \\ p_{poiss}(n_{\mathcal{A}}(y)|\lambda_A) & \text{if } y > \mathcal{A}.y_{grad} \end{cases}$$

<sup>2</sup>In case  $\mathcal{A}.y_{grad}$  or  $\mathcal{A}.y_{end}$  lie beyond the last year of publications recorded in DBLP (i.e. the current year,  $y_{current} = 2007$ ) the probability distribution  $p(n_{\mathcal{A}}(y)|\mathcal{H}, \Lambda)$  becomes identically zero for all  $y > y_{current}$

Finally, every author  $\mathcal{A}$  chooses his co-authors for a publication  $\mathcal{P}$  in year  $y$  based on his academic connections during the year  $y$ . The probability distribution is modeled differently based on whether  $\mathcal{A}$  is a student or has already graduated in year  $y$  (i.e. based on whether  $y \leq \mathcal{A}.y_{grad}$  or not). Let  $\mathcal{P} = (\mathcal{C}, y_{\mathcal{P}})$  where,  $\mathcal{C} = \{\mathcal{A}_{\mathcal{P}}(1), \mathcal{A}_{\mathcal{P}}(2) \dots \mathcal{A}_{\mathcal{P}}(k_{\mathcal{P}})\}$  is the list of authors in  $\mathcal{P}$ .

1. If  $\mathcal{A}$  is a student during year  $y$ ,  $\mathcal{A}$  chooses his advisor  $\mathcal{A}.adv$  as a co-author with probability  $p_{adv}$ . Let  $G(\mathcal{A}, y)$  denote the set containing all the students of  $\mathcal{A}.adv$  in year  $y$ . The student also chooses each of the other group members  $g \in G(\mathcal{A}, y)$  in the students group in year  $y$  as co-authors with probability  $p_{s,grp}$ . Finally, the author picks co-authors from any of the remaining active authors in the database again with an equal (extremely small) probability. Since the number of such possible co-authors is extremely large, and each of them is chosen with an extremely small probability we may approximate the number of co-authors chosen using the Poisson Distribution. Hence we assume that the mean number of additional (out of group) co-authors is generated from a Poisson distribution with mean  $\lambda_{s,c}$ . The actual co-authors are allowed to be any of the remaining authors (who are also active during year  $y$ ) in the DBLP database with an equal probability. For the sake of simplicity we approximate this probability by  $1/n_{db}$ , where  $n_{db}$  (representing the number of active authors in year  $y$ ) is chosen to be a constant.

Let  $Z$  be the indicator variable that is 1 or 0 based on whether the advisor of the first author belongs to the author list or not (i.e. whether  $\mathcal{A}_{\mathcal{P}}(1).adv \in \mathcal{C}$  or not). Let  $Z^g$  be similar indicator variables for each of the group members  $g \in G(\mathcal{A}_{\mathcal{P}}(1), y)$ . That is,  $Z^g$  is either 1 or 0 based on whether  $g \in \mathcal{C}$  or not. Finally, let the number of external (neither the advisor, nor part of the group) co-authors in  $\mathcal{P}$  be  $n_{ext}(\mathcal{P})$ . Hence, if  $\mathcal{A}_{\mathcal{P}}(1)$  is a student during year  $y$ , we compute  $p(\mathcal{P}|\mathcal{H}, \Lambda)$  as,

$$\begin{aligned} p(\mathcal{P}|\mathcal{H}, \Lambda) &= p_{adv}^Z (1 - p_{adv})^{(1-Z)} \\ &\times \prod_{g \in G(\mathcal{A}_{\mathcal{P}}(1), y) \setminus \mathcal{A}_{\mathcal{P}}(1)} p_{s,grp}^{Z^g} (1 - p_{s,grp})^{(1-Z^g)} \\ &\times p_{poiss}(n_{ext}(\mathcal{P})|\lambda_{s,c}) \left( \frac{1}{n_{db}} \right)^{n_{ext}(\mathcal{P})} \end{aligned}$$

2. If  $\mathcal{A}$  has already graduated, he generates his co-authors as follows: Again we define  $G(\mathcal{A}, y)$  as the set of all the students in the research group of  $\mathcal{A}$  during year  $y$ . If  $\mathcal{A}$  has already graduated this is the set of students of  $\mathcal{A}$  during year  $y$ .  $\mathcal{A}$  chooses each of his group members as co-authors for his publication with probability  $p_{a,grp}$ . Finally as in the previous case, we assume that the author chooses the number of other co-authors in  $\mathcal{P}$  from a Poisson distribution with mean  $\lambda_{a,c}$ . The actual co-authors are

allowed to be any of the remaining authors (who are also active during year  $y$ ) in the DBLP database with an equal probability which is again approximated as  $1/n_{db}$ .

Again, we define the indicator variables  $Z^g$  to be either 1 or 0 based on whether  $g \in \mathcal{C}$  or not. Also, let the set of external (neither the advisor, nor part of the group) co-authors in  $\mathcal{P}$  be  $n_{ext}(\mathcal{P})$ . Hence, if the author  $\mathcal{A}_{\mathcal{P}}(1)$  has already graduated during year  $y$ , we compute  $p(\mathcal{P}|\mathcal{H}, \Lambda)$  as,

$$p(\mathcal{P}|\mathcal{H}, \Lambda) = \prod_{g \in G(\mathcal{A}_{\mathcal{P}}(1), y) \setminus \mathcal{A}_{\mathcal{P}}(1)} p_{a,grp}^{Z^g} (1 - p_{a,grp})^{(1-Z^g)} \\ \times p_{poiss}(n_{ext}(\mathcal{P})|\lambda_{a,c}) \left( \frac{1}{n_{db}} \right)^{n_{ext}(\mathcal{P})}$$

## 4 Model Learning

The previous section described the generative model  $p(\mathcal{H}, \mathcal{V}|\Lambda)$  in terms of the observation model  $p(\mathcal{V}|\mathcal{H}, \Lambda)$  and the prior probability model  $p(\mathcal{H}|\Lambda)$ . As mentioned in Section 2, are interested in learning the parameters  $\Lambda$  and then inferring the maximum-a-posteriori (or MAP) estimate of the hidden variables. The standard method for learning the parameters of such a generative model is the Expectation-Maximization (or EM) algorithm (see [2] for a tutorial).

The algorithm is initialized with some initial estimate for the parameters  $\Lambda$  and proceeds by iterating two steps:

- **E Step:** In the expectation step, the algorithm computes the expected value of the complete-data log-likelihood  $\log p(\mathcal{H}, \mathcal{V}|\Lambda)$  with respect to the unknown hidden variables  $\mathcal{H}$  given the observed data  $\mathcal{V}$  and the current parameter estimates. That is we define:

$$q^i(\mathcal{H}) = p(\mathcal{H}|\mathcal{V}, \Lambda^i) \quad (3)$$

$$\mathcal{L}(\Lambda, \Lambda^i) = E_{q^i} [\log p(\mathcal{H}, \mathcal{V}|\Lambda)] \quad (4)$$

Here,  $q^i(\mathcal{H})$  is the posterior distribution over the hidden variables  $\mathcal{H}$  given the parameter estimates  $\Lambda^i$  after the  $i^{th}$  iteration and  $\mathcal{L}(\Lambda, \Lambda^i)$  is the expectation under  $q^i(\mathcal{H})$  of the complete data log-likelihood using different parameters  $\Lambda$ .

- **M Step:** In the maximization step, the algorithm maximizes the computed expectation  $\mathcal{L}(\Lambda, \Lambda^i)$  to find the new parameter estimates  $\Lambda^{i+1}$ . That is:

$$\Lambda^{i+1} = \arg \max_{\Lambda} \mathcal{L}(\Lambda, \Lambda^i)$$

In our case, computing the exact posterior distribution  $q(\mathcal{H})$  in the  $E$ -step is not feasible since the posterior distribution is over an extremely high dimensional space<sup>3</sup>. Also, the posterior cannot be factored into a product of distributions over smaller independent subsets due to dependency among co-authors. In such a case, a standard solution is to approximate the true posterior  $p(\mathcal{H}|\mathcal{V}, \Lambda)$  with a variational approximation  $q(\mathcal{H})$  [4] where the functional form of the approximation can be factorized in a simple form. In our case, we may assume that  $q(\mathcal{H})$  can be factored as a product of marginal distributions for the different authors. That is,  $q(\mathcal{H}) = \prod_{\mathcal{A} \in DBLP} q(\mathcal{H}_{\mathcal{A}})$  where  $q(\mathcal{H}_{\mathcal{A}})$  is the marginal distribution over the hidden variables corresponding to author  $\mathcal{A}$ . This is referred to as the *mean field approximation* [4]. In the variational EM setting, we look to find the variational distribution  $q(\mathcal{H})$  that minimizes the Kullback-Leibler (or KL) Divergence to the true posterior. That is, we are interested in the  $q(\mathcal{H})$  that minimizes  $KL(q(\mathcal{H})||p(\mathcal{H}|\mathcal{V}, \Lambda))$ . The KL Divergence [7] is defined as:

$$KL(q(\mathcal{H})||p(\mathcal{H}|\mathcal{V}, \Lambda)) = \sum_{\mathcal{H}} q(\mathcal{H}) \log \frac{q(\mathcal{H})}{p(\mathcal{H}|\mathcal{V}, \Lambda)}$$

In our case, the distribution  $q(\mathcal{H})$  defined using the mean field approximation is still intractable to compute due to the huge dimensionality of the hidden variable space. Hence, we take the variational approximation one step further and approximate each  $q(\mathcal{H}_{\mathcal{A}})$  using a single delta function. That is, we force  $\mathcal{H}_{\mathcal{A}}$  to take on one specific value with probability 1 and all other values with probability zero. In this case,  $q(\mathcal{H})$  must also be a single delta function and finding the  $q(\mathcal{H})$  that minimizes the KL divergence with the true posterior  $p(\mathcal{H}|\mathcal{V}, \Lambda)$  amounts to simply finding the mode of the true posterior distribution  $p(\mathcal{H}|\mathcal{V}, \Lambda)$ . Mathematically, using this variational approximation amounts to simply replacing the exact definition of  $q^i(\mathcal{H})$  in equation 3 with the following approximation:

$$\begin{aligned} \mathcal{H}_i^* &= \arg \max_{\mathcal{H}} p(\mathcal{H}|\mathcal{V}, \Lambda^i) \\ q(\mathcal{H}) &= \begin{cases} 1 & \text{if } \mathcal{H} = \mathcal{H}_i^* \\ 0 & \text{o/w} \end{cases} \end{aligned} \quad (5)$$

Under some technical conditions (which are almost always satisfied), the EM-algorithm is guaranteed to converge to parameter estimates that are a local maximum of likelihood  $p(\mathcal{V}|\Lambda) = \sum_{\mathcal{H}} p(\mathcal{H}, \mathcal{V}|\Lambda)$  of the observed variables  $\mathcal{V}$ . This is exactly the function we had set out to optimize at the start (equation 1). However, this nice convergence guarantee does not hold in our variational EM setting. Nevertheless, since the expectation computed after the  $E$ -step (which is the complete data log likelihood in our case) always increases after every iteration and is bounded above, the variational EM iterations also converge. The quality of the resulting parameter estimates depends on how good the

---

<sup>3</sup>The number of hidden variables is  $4N_{DBLP}$ , where  $N_{DBLP}$  is the number of authors in DBLP

variational approximation is in practice. In our case, since we expect the true marginal posterior distributions  $p(\mathcal{H}_{\mathcal{A}}|\mathcal{V}, \Lambda)$  to be highly peaked (the probability of the correct advisor to be much larger than the other options), we expect the procedure to converge to reasonable parameter estimates. The quality of the parameter estimates may be further increased by using say the top  $k$  possible values for  $p(\mathcal{H}_{\mathcal{A}}|\mathcal{V}, \Lambda)$  instead of just the mode. Looking at other efficiently computable approximations is left for future research.

Inference becomes trivial under the approximation described in equation 5. Recall (from equation 2) that we are interested in searching for the MAP estimate  $\mathcal{H}^*$  after learning the parameters. In other words, we are searching for the mode of  $p(\mathcal{H}|\mathcal{V}, \Lambda)$  which is exactly what we compute in the E-step of the variational EM algorithm. Hence, we may run the E-step once more to finally compute the mode  $\mathcal{H}^*$ .

#### 4.1 E Step: Computing the Mode of $p(\mathcal{H}|\mathcal{V}, \Lambda)$

Since computing the exact mode is not feasible due to the high dimensionality of  $\mathcal{H}$ , we use an iterative algorithm to find a local maximum of  $p(\mathcal{H}|\mathcal{V}, \Lambda)$ . The new mode is initialized with the estimate of the hidden variables obtained at the end of the previous variational EM iteration. Next, the hidden variables  $\mathcal{H}_{\mathcal{A}} = (\mathcal{A}.adv, \mathcal{A}.y_{start}, \mathcal{A}.y_{grad}, \mathcal{A}.y_{end})$  for each author in the database are updated to their optimal values keeping all the variables corresponding to the other authors “fixed”<sup>4</sup>.

Since, each step in this iteration increases the complete data log likelihood  $p(\mathcal{H}, \mathcal{V}|\Lambda)$ , these iterations must converge as well. However, due to the high computational cost of each of these iterations (and because the gains in subsequent iterations become small), only 1 or 2 of these iterations are used to update the hidden variables after every iteration of variational EM.

#### 4.2 M Step: Updating the Parameters

In the maximization step of the variational EM algorithm, we need to update the parameters  $\Lambda^i$ . This process can be accomplished efficiently due to the simple choice of the component prior and observation model distributions as well as the simple variational approximation chosen. We know that the maximum likelihood value of the mean parameter for any Poisson distribution is just the mean of the observed samples from that distribution. Let  $\mathcal{P}.fa$  denote the first author in publication  $\mathcal{P}$  and  $\mathcal{P}.y$  denote the year of publication of  $\mathcal{P}$ . Also, let  $n_{grp}(\mathcal{P})$  denote the number of group members of the first author in the author list. We may update the Poisson parameters as:

---

<sup>4</sup>Note that some configurations may generate inconsistencies: for example when  $\mathcal{A}$ 's graduation year is increased some of his students may now start working under him before this new graduation year which is not allowed by the model. In such cases alternative advisors are found for such students. On the other hand, when  $\mathcal{A}$ 's graduation year is decreased, other authors who could not be his students earlier can now be his students and these options are also tested while computing the best values for the hidden variables  $\mathcal{H}_{\mathcal{A}}$ .

$$\begin{aligned}\lambda_N &\leftarrow \text{avg}_{\mathcal{A}} \{(\mathcal{A}.y_{grad} + 1) - \mathcal{A}.y_{start}\} \\ \lambda_{act} &\leftarrow \text{avg}_{\mathcal{A}} \{\mathcal{A}.y_{end} - \mathcal{A}.y_{grad}\} \\ \lambda_{s,c} &\leftarrow \text{avg}_{\mathcal{P}, isStudent(\mathcal{P}.fa)} n_{ext}(\mathcal{P}) \\ \lambda_{a,c} &\leftarrow \text{avg}_{\mathcal{P}, hasGraduated(\mathcal{P}.fa)} n_{ext}(\mathcal{P})\end{aligned}$$

Similarly, the probability parameters may also be updated as:

$$\begin{aligned}p_{s,grp} &\leftarrow \frac{\sum_{\mathcal{P}, isStudent(\mathcal{P}.fa)} n_{grp}(\mathcal{P})}{\sum_{\mathcal{P}, isStudent(\mathcal{P}.fa)} |G(\mathcal{P}.fa, \mathcal{P}.y) - 1|} \\ p_{a,grp} &\leftarrow \frac{\sum_{\mathcal{P}, hasGraduated(\mathcal{P}.fa)} n_{grp}(\mathcal{P})}{\sum_{\mathcal{P}, hasGraduated(\mathcal{P}.fa)} |G(\mathcal{P}.fa, \mathcal{P}.y) - 1|} \\ p_{db} &\leftarrow \frac{\#(\mathcal{A}.adv \in DBLP)}{N_{DBLP}} \\ p_{adv} &\leftarrow \frac{\#(isStudent(\mathcal{P}.fa) \ \&\& \ \mathcal{P}.fa.adv \in \mathcal{P})}{\#(isStudent(\mathcal{P}.fa))}\end{aligned}$$

## 5 Implementation Results

**Initialization:** Before starting the variational EM algorithm, the parameters are initialized to reasonable values. The values used are

$$\begin{aligned}\lambda_N &= 5 \\ \lambda_{act} &= 10 \\ \lambda_{s,c} &= 1 \\ \lambda_{a,c} &= 2 \\ p_{s,grp} &= 0.1 \\ p_{a,grp} &= 0.25 \\ p_{db} &= 0.8 \\ p_{adv} &= 0.5\end{aligned}$$

The values for  $\lambda_A$  and  $\lambda_S$  are initialized as follows. We assume that all the authors with less than or equal to 3 publications correspond to students. The mean number of publications generated by these students per year during the time between the first and last publications is used to initialize  $\lambda_S$ . Next, we assume that all the authors with greater than or equal to 15 years of activity are all advisors during the last 3 years of activity. The mean number of first

author publications generated by them per year during these years are used to estimate  $\lambda_A$ . The values turn out to be:

$$\begin{aligned}\lambda_A &= 0.0156453 \\ \lambda_S &= 0.415374\end{aligned}$$

After only 5 iterations (which is 10 E and M steps) of the variational EM method the parameters stabilize to

$$\begin{aligned}\lambda_A &= 0.056538 \\ \lambda_S &= 0.258349 \\ \lambda_N &= 2.71154 \\ \lambda_{Act} &= 10.276 \\ \lambda_{s,c} &= 0.601365 \\ \lambda_{a,c} &= 1.20665 \\ p_{db} &= 0.66483 \\ p_{adv} &= 0.849647 \\ p_{s,grp} &= 0.117341 \\ p_{a,grp} &= 0.268831\end{aligned}$$

All the values converge to reasonable estimates. The only interesting point to notice is that  $\lambda_N$  converges to less than 3 years. There are several reasons for this effect: 1) there are authors that do not complete a PhD and graduate with some other degree (for example a Masters), 2) some PhD students are not active during the first few years of their PhD and 3) many institutes in Europe have average graduation times of about 3 years. Figure 2 shows the negative complete data log likelihood during these 5 EM iterations.

For quantitative performance evaluation, a labeled test set was generated consisting of 166 PhD students (old and current) of 10 different professors in the Computer Science Department at the University of Illinois, Urbana-Champaign. Only, 114 of these students were present in the DBLP database. This happened because 1) some of the students were new and had not yet published and 2) the names of some of the students on the websites did not exactly match those in DBLP. For estimating the hardness of the test set, a baseline approach was implemented. In the baseline method, the dates  $y_{start}$ ,  $y_{grad}$  and  $y_{end}$  were fixed for all the authors using simple heuristics.  $y_{start}$  was assigned to be the first year of publication.  $y_{grad}$  was assigned to be  $y_{start} + \lambda_N$  ( $y_{start} + 5$ ). Finally,  $y_{end}$  was assigned to be the last publication year. The advisor for every author was chosen to be the most frequent graduated co-author. In case of a tie, the publications with lesser number of authors were given more weight. The baseline method gets 68 out of the 114 advisors correct giving an accuracy of 59.6%. The Variational EM method gets 80 out of the 114 advisors correct

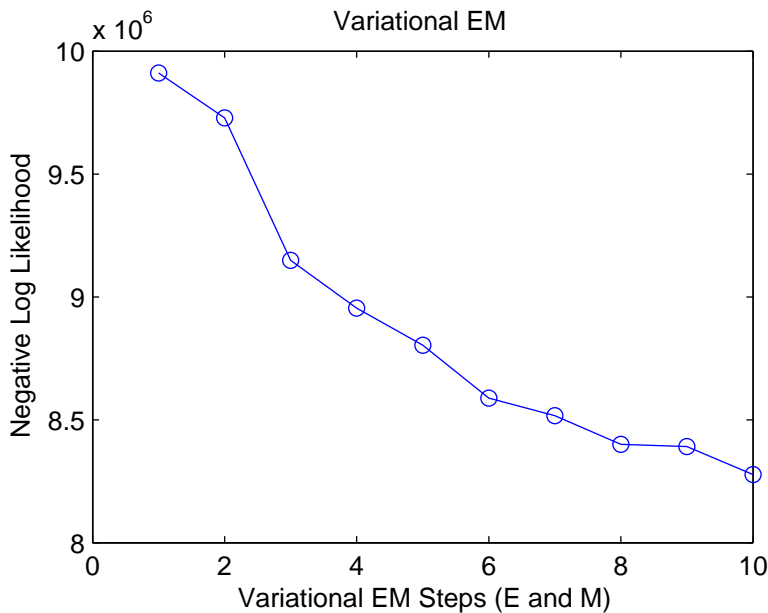


Figure 2: Plot of the Negative Complete Data Log Likelihood during the Variational EM

giving an accuracy of 70.2%. There is a substantial increase in the number of correctly labeled advisors in the case of variational EM. Another advantage of using the variational EM method is that one can estimate the approximate graduation years as a by-product of the computation. Also, the variational EM method can be extended to systematically define and find collaborators.

An analysis of the errors made by the variational EM algorithm revealed that quite a few of them happened on students with very common names that have entries for many other students clubbed into their entry in DBLP. It might be beneficial to employ an approach like that of Yin et al. [10] to separate out the entries corresponding to different authors with similar names as a preprocessing step.

### 5.1 Demo:

The advisor and graduation information extracted has been put into a database that can be queried using a PHP script at:

<http://www-cvr.ai.uiuc.edu/~kushal/courses/cs512/advisor.php>

## 6 Conclusions and Future Work

A probabilistic generative model was proposed to extract advisor-advisee relationships from a bibliography database. The approach was implemented and

shown to perform well on the DBLP database. At this point there are multiple directions in which this work can be improved or extended.

- The variational approximation used during the EM procedure can be made more accurate. One of the ideas (also mentioned earlier) is to use the top  $k$  values of the variables instead of just the mode while defining  $q(\mathcal{H}_{\mathcal{A}})$ .
- The model itself can be extended in various ways to improve its modeling accuracy.
  1. Since, some students may switch advisors during the first few years of their research career one may want to allow  $\mathcal{A}.y_{start}$  to occur sometime after the first few publications. One more parameter that controls this amount of deviation from the first publication would need to be introduced in this case.
  2. One may also want to add collaborator relationships into the model. We could fix a bound on the maximum number of collaborators for an author or have a distribution on the number of collaborators based on the number of years of active research for the author. Collaborator relations could then be modeled similar to advisor and group relations. A probability parameter  $p_{col}$  could be introduced to denote the probability that a collaborator of  $\mathcal{A}$  during a specific year would be included in the author list of the publications generated by  $\mathcal{A}$  during that year.
- One may preprocess the database to separate distinct authors with identical names as well as combine the different names corresponding to the same author based on their common co-authors (similar to [10], [1]). Doing this correctly can result in a reasonable gain in performance as many errors are caused due to such cases.

## References

- [1] Indrajit Bhattacharya and Lise Getoor. Iterative record linkage for cleaning and integration. In *The SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD)*, Paris, France, 2004.
- [2] J. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.*, 1997.
- [3] Harry B. Coonce. The Mathematics Genealogy Project <http://www.genealogy.ams.org/>.
- [4] T. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice. MIT Press*, 2000.

- [5] Benjamin Kuipers. AI Genealogy Project.  
<http://aigp.csres.utexas.edu/~aigp/>.
- [6] Michael Ley. DBLP: Computer Science Bibliography Database  
<http://www.sigmod.org/dblp/db/>.
- [7] Wikipedia. Kullback Leibler divergence.
- [8] Tao Xie. The Software Engineering Academic Genealogy  
<http://www.csc.ncsu.edu/faculty/xie/sefamily.htm>.
- [9] Yuan Xie. Computer Engineering Academic Genealogy  
<http://www.cse.psu.edu/~yuanxie/community/genealogy/>.
- [10] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. *ICDE*, 2007.